Proceedings of the COLING 2004 Post Conference Workshop on Multilingual Linguistic Ressources MLR2004

Gilles Sérasset, Susan Armstrong, Christian Boitet, Andrei Popescu-Belis, Dan Tufis

COLING 2004

The 20th International Conference on Computational Linguistics
Post-Conference Workshop

Proceeding of the Workshop on Multilingual Linguistic Ressources MLR2004

Editors:
Gilles Sérasset, Susan Armstrong, Christian Boitet,
Andrei Popescu-Belis, Dan Tufis

August 28th, 2004
University of Geneva, Switzerland

Foreword

In an ever expanding information society, most information systems are now facing the "multilingual challenge". Multilingual language resources play an essential role in modern information systems. Such resources need to provide information on many languages in a common framework and should be (re)usable in many applications (for automatic or human use).

Many centres have been involved in national and international projects dedicated to building harmonised language resources and creating expertise in the maintenance and further development of standardised linguistic data. These resources include dictionaries, lexicons, thesauri, word-nets, and annotated corpora developed along the lines of best practices and recommendations. However, since the late 90's, most efforts in scaling up these resources remain the responsibility of the local authorities, usually, with very low funding (if any) and few opportunities for academic recognition of this work. Hence, it is not surprising that many of the resource holders and developers have become reluctant to give free access to the latest versions of their resources, and their actual status is therefore currently rather unclear.

The goal of this workshop is to study problems involved in the development, management and reuse of lexical resources in a multilingual context. Moreover, this workshop provides a forum for reviewing the present state of language resources. The workshop is meant to bring to the international community qualitative and quantitative information about the most recent developments in the area of linguistic resources and their use in applications.

The impressive number of submissions (38) to this workshop and in other workshops and conferences dedicated to similar topics proves that dealing with multilingual linguistic ressources has become a very hot problem in the Natural Language Processing community.

To cope with the number of submissions, the workshop organising committee decided to accept 16 papers from 10 countries based on the reviewers' recommendations. Six of these papers will be presented in a poster session. The papers constitute a representative selection of current trends in research on Multilingual Language Resources, such as multilingual aligned corpora, bilingual and multilingual lexicons, and multilingual speech resources. The papers also represent a characteristic set of approaches to the development of multilingual language resources, such as automatic extraction of information from corpora, combination and re-use of existing resources, online collaborative development of multilingual lexicons, and use of the Web as a multilingual language resource.

The development and management of multilingual language resources is a long-term activity in which collaboration among researchers is essential. We hope that this workshop will gather many researchers involved in such developments and will give them the opportunity to discuss, exchange, compare their approaches and strengthen their collaborations in the field.

The organisation of this workshop would have been impossible without the hard work of the program committee who managed to provide accurate reviews on time, on a rather tight schedule. We would also like to thank the Coling 2004 organising committee that made this workshop possible. Finally, we hope that this workshop will yield fruitful results for all participants.

Gilles Sérasset
Organising chair
GETA (Study Group for Machine Translation), CLIPS-IMAG laboratory
Université Joseph Fourier, France

Program Committee

Gilles Sérasset *(Chair)*	GETA CLIPS-IMAG, Université Joseph Fourier - Grenoble I, France
Susan Armstrong	ISSCO, Université de Genève, Switzerland
Pushpak Battacharya	IIT, Mumbai, India
Igor Boguslavski	IITP, Moscow, Russia
Christian Boitet	GETA CLIPS-IMAG, Université Joseph Fourier - Grenoble I, France
Pierrette Bouillon	ISSCO, Université de Genève, Switzerland
Jim Breen	Monash University, Australia
Nicoletta Calzolari	CNR, Pisa, Italy
Dan Cristea	University Al.I.Cuza Iasi, Romania
Patrick Drouin	OLST, University of Montreal,Canada
Sanae Fujita	NTT, Kyoto, Japan
Ulrich Heid	IMS-CL, University of Stuttgart, Germany
Hitoshi Isahara	CRL, Nara, Japan
Kyo Kageura	NII, Tokyo, Japan
Chuah Choy Kim	USM, Penang, Malaisie
Mathieu Mangeot	NII, Tokyo, Japan
Alain Polguère	OLST, University of Montreal,Canada
Andrei Popescu-belis	ISSCO, Université de Genève, Switzerland
Jean Senellart	SYSTRAN, France
Mandel Shi	Xiamen University, China
Virach Sornlertlamvanich	Thai Computational Linguistics Laboratory, CRL, Thailand
Pr. Kumiko Tanaka-Ishii	Tokyo University, Japan
Philippe Thoiron	CRTT, Université de Lyon 2, France
Dan Tufis	RACAI, Uni Bucharest, Romania
Michael Zock	LIMSI, Orsay, France

Tentative Program

Time	Event
08:30 — 09:00	Registration & Welcome
09:00 — 10:00	Paper Session • **JMdict: a Japanese-Multilingual Dictionary** • **A Generic Collaborative Platform for Multilingual Lexical Databases Development**
10:00 — 11:00	Poster Session and opened discussions • **Semi-Automatic Construction of Korean-Chinese Verb Patterns based on Translation Equivalency** • **Bilingual Sign Language Dictionary to Learn the Second Sign Language without Learning a Target Spoken Language** • **Building Parallel Corpora for eContent Professionals** • **Revising the WordNet Domains Hierarchy: semantics, coverage and balancing** • **PolyphraZ: a tool for the management of parallel corpora** • **Multilingual Text Induced Spelling Correction**
11:00 — 11:30	Coffee Break
11:30 — 13:00	Papers Session • **A Model for Fine-Grained Alignment of Multilingual Texts** • **Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora** • **Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications**
13:00 — 14:00	Lunch Break

Time	**Event**
14:00 — 15:00	Papers Session • **A Method of Creating New Bilingual Valency Entries using Alternations** • **Automatic Construction of a Transfer Dictionary Considering Directionality**
15:00 — 15:30	Poster Session and opened discussions • **Semi-Automatic Construction of Korean-Chinese Verb Patterns based on Translation Equivalency** • **Bilingual Sign Language Dictionary to Learn the Second Sign Language without Learning a Target Spoken Language** • **Building Parallel Corpora for eContent Professionals** • **Revising the WordNet Domains Hierarchy: semantics, coverage and balancing** • **PolyphraZ: a tool for the management of parallel corpora** • **Multilingual Text Induced Spelling Correction**
15:30 — 16:00	Coffee Break
16:00 — 17:30	Papers Session • **Building and sharing multilingual speech resources, using ERIM generic platforms** • **Multilinguality in ETAP-3: Reuse of Lexical Resources** • **Qualitative Evaluation of Automatically Calculated Acception Based MLDB**
17:30 — 18:00	Opened discussions & Closing

Contents

Multilinguality in ETAP-3: Reuse of Lexical Resources

Igor BOGUSLAVSKY
Universidad Politecnica de Madrid
28660 Boadilla del Monte, Madrid, Spain
igor@opera.dia.fi.upm.es

Leonid IOMDIN
Institute for Information Transmission Problems, Russian Academy of Sciences
19, B. Karetnyj
Moscow, GSP-4, Russia
iomdin@cl.iitp.ru

Victor SIZOV
Institute for Information Transmission Problems, Russian Academy of Sciences
19, B. Karetnyj
Moscow, GSP-4, Russia
sizov@cl.iitp.ru

Abstract

The paper presents the work done at the Institute for Information Transmission Problems (Russian Academy of Sciences, Moscow) on the multifunctional linguistic processor ETAP-3. Its two multilingual options are discussed – machine translation in a variety of language pairs and translation to and from UNL, a meaning representation language.

For each working language, ETAP has one integral dictionary, which is used in all applications both for the analysis and synthesis (generation) of the given language. In difficult cases, interactive dialogue with the user is used for disambiguation. Emphasis is laid on multiple use of lexical resources in the multilingual environment.

1 General Information on ETAP

The multifunctional ETAP-3 linguistic processor, developed by the Computational Linguistics Laboratory (CLL) in Moscow (see e.g. Apresjan *et al.* 1992a,b, 1993, 2003), is the product of more than two decades of laboratory research and development in the field of language modeling. The most important features of the processor are as follows.

(1) ETAP-3 is based on the **general linguistic framework** of the Meaning ⇔ Text theory, proposed by Igor Mel'cuk (e.g. Mel'cuk, 1974) and complemented by the theory of systematic lexicography and integrated description of language proposed by Jurij Apresjan [Apresjan 1995, 2000].

(2) ETAP-3 has a declarative organization of linguistic knowledge.

(3) One of the major components of ETAP-3 is the innovative combinatorial dictionary. Apart from syntactic and semantic features and subcategorization frames, the dictionary entry may have rules of 8 types. Many dictionary entries contain **lexical functions** (LF).

(3) ETAP-3 makes use of a formalism based on three-value predicate logic, in which all linguistic data are presented.

(4) The ETAP-3 processor has **a modular architecture**. All stages of processing and all types of linguistic data are organized into modules, which warrants their reusability in many NLP applications both within and beyond ETAP-3 environment.

At the moment, the ETAP-3 environment comprises the following main options: 1) a rule-based machine translation system; 2) a Universal Networking Language (UNL) translation engine; 3) a system of synonymous paraphrasing of sentences; 4) a workbench for syntactic annotation of text corpora; and 5) a grammar checker. All the applications make use of the same dictionaries, but only the first and the second are multilingual. In Section 2 we will discuss multilingual lexical resources used in machine translation, and in Section 3 – in the UNL module.

2 Multilinguality in ETAP

2.1 Structure of the Dictionary Entry

To support multilinguality, the dictionary entry of the ETAP dictionary has several sub-zones. There is one general zone and several zones oriented towards various languages. The general zone stores all types of monolingual information: part of speech, syntactic features, semantic features, subcategorization frames, lexical functions, syntactic and pre-syntactic rules, generation rules, and some other data. Each bi-lingual sub-zone serves for establishing

correspondence between the given language and another one (see Fig. 1).

For example, the Russian zone of an English dictionary entry contains all the information needed to translate English words into Russian, the Arabic zone provides translation into Arabic, etc. Conversely, the information needed to translate Russian words into English is stored in the English zone of the Russian dictionary entries.

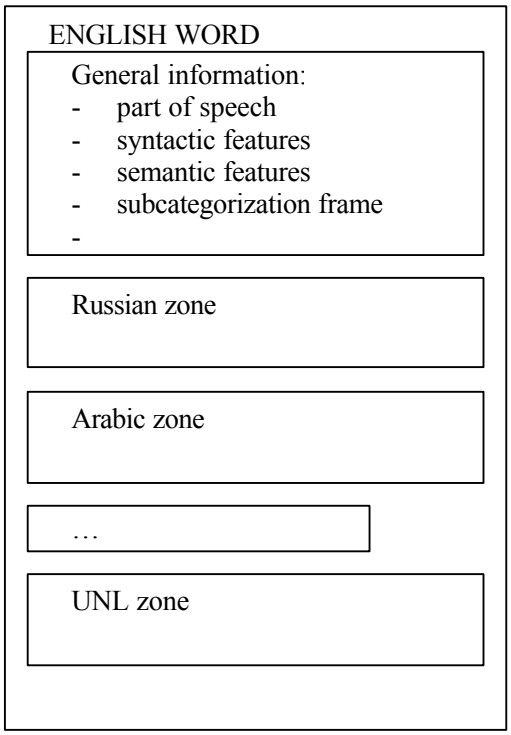


Fig. 1

2.2 Default and Specific Translation

The information stored in a bi-lingual zone consists of two parts: a default translation and lexical translation rules. Default translation is a single word that translates the given word in non-specific contexts (it is introduced by a special label: TRANS). Any other type of translation is carried out by means of rules. If the word is translated by a phrase consisting of several words, the rule shows how the words in the phrase are connected to each other and how this phrase is incorporated into the sentence. For example, in the entry *bachelorship* we find a reference to one of the standard translation rules (TRADUCT2.42). The slots of the rule are filled with specific lexical items, grammatical features or syntactic relations.

TRAF:TRADUCT2.42
LR1:STEPEN',LR2:BAKALAVR,T2:SG,
T3:QUASIAGENT

The rule says that *bachelorship* should be translated into Russian with a phrase consisting of two words – *stepen'* ('degree') and *bakalavr* ('bachelor'). These words should be connected by the quasiagent(ive) syntactic relation, and the number feature of *bakalavr* should be singular.

If the word is translated in a specific way in a specific context or in specific phrases, the rule describes this context and the resulting structure. When a word is translated, normally first the translation rules in its dictionary entry are tried. If no rule applies in the given sentence, then the default translation is used.

2.3 Multiple Translation

The default option of ETAP produces a single translation of the sentence – the one that corresponds to the first lexico-syntactic structure obtained by the parser. The option of multiple translation produces much more. First, it generates all lexico-syntactic structures that are compatible with the grammar and the dictionary. Since these structures are disambiguated both syntactically, and lexically, this set of structures contains all lexical variants for the source sentence. Then, for each structure all possible translation variants are tried. As is known, even disambiguated words can be translated into another language in different ways and it is not always possible to formulate a rule that could select an appropriate variant. For example, English *adjuration* can be translated into Russian as *mol'ba* and as *zaklinanie, adventurer* – as *avantjurist* and as *iskatel' prikljuchenij* (literally, 'adventure seeker'), *alarm* – as *trevoga* and as *avarijnyj signal* ('alarm signal'). In all these cases, we are most probably dealing with a single meaning of the English word and yet translation variants are not fully synonymous. Since we cannot choose among these variants by means of rules and at the same time do not want to lose any of them, we have to treat them as alternative translations to be activated in the "Multiple translation" option. As mentioned in the previous section, there are two types of translation devices in the bilingual zones of the dictionary: a default translation (a single word) and rules. In both cases, it is possible to provide alternative translations. For example, in the entry for *adjuration* alternative translations are listed in the default part since both of them are single words:

ADJURATION
…
TRANS: *MOL'BA / ZAKLINANIE*

If the user selects the "Single translation" option, only the first of these variants will be used. If

he/she wishes to get all possible translations and activates the "Multiple translation" option, both alternatives will be produced.
In the *adventurer* entry, the alternative translation *iskatel' prikljuchenij* should be introduced by a rule, since it is not a single word but a phrase. Such rules are supplied by a special marker, OPT(ional), which shows that the translation is alternative.

ADVENTURER
…
TRANS: *AVANTJURIST*
TRAF:TRADUCT2.42
OPT:1
LR1:*ISKATEL'2*,LR2:*PRIKLJUCHENIE*,T2:PL,
T3: ATTRIB

This is another instance of the same rule that we saw above in the *bachelorship* example: the only difference is that it introduces different words, connects them with a different syntactic relation (attributive) and generates a different number feature. The marker OPT:1 shows that the translation introduced by this rule is less common than the default translation *avantjurist* and should be presented to the user after it. Should it be otherwise, the rule would have the marker OPT:0 and have a priority over the default translation.

2.4 Interactive selection of the translation equivalent

It is well known that ambiguity of linguistic units is one of the most difficult problems in NLP. In ETAP there is no single stage of processing that expressly deals with disambiguation. The sentence is gradually disambiguated at different stages of processing on the basis of restrictions imposed by the linguistic knowledge of the system. However, in many cases this knowledge is not sufficient for complete disambiguation, since the understanding of a text by humans is not based on their linguistic knowledge alone. To cope with this problem, we are developing an interactive option that at certain pivotal points of text processing is expected to ask for human intervention and use human assistance to resolve those ambiguities that are beyond the scope of linguistic knowledge of the system (Boguslavsky et al 2003). It should be stressed that the interactive tool is only resorted to if an ambiguity cannot be resolved automatically and therefore requires human intervention. This work is in line with the approach proposed in a series of publications by the GETA group (Blanchon, 1995, 1996, 1997, Boitet & Blanchon, 1995).

As mentioned above, the dialogue with the user is activated at different stages of the processing depending on the tasks solved at each stage. During the parsing, which results in the construction of the lexico-syntactic structure of the sentence, all lexical and syntactic ambiguity should be resolved. However, this is done entirely within the processing of the source language text and represents monolingual ambiguity. It is not directly relevant for our topic of multilinguality. Of relevance here are cases of the so-called translational (or transfer) ambiguity (Hutchins, Somers, 1992: 87). The source language words can be unambiguous for the native speakers of this language but can be translated by a number of different target language expressions. In this sense, they are ambiguous from the viewpoint of the target language and have to be dealt with at the translation stage. An example is the English verb *wash* with respect to Russian. It translates differently depending on the type of object that is being washed: if it is something made of cloth, for example clothes, a special verb has to be chosen. If the dictionary provides semantic information on what objects are made of, the correct choice of the verb can in principle be made automatically. Cf., however, cases like *We must wash it* where such information is definitely missing.

This must be viewed as a relatively inoffensive case, though, because most sentences will be translated correctly with the help of a simple rule (and if not, the mistake is not too important). There are many words for which it is much more difficult to write a disambiguation rule. A notorious example is English *blue* that corresponds to two Russian adjectives, one meaning 'light blue' and the other – roughly – 'dark blue'. The only way to translate this word correctly in most of the contexts is to get assistance from the user. The dialog with the user is based on the information stored in the dictionary and activated at the appropriate moment.

This is how the interactive disambiguation currently works. The sentence to be translated is entered in the upper window of the ETAP environment (Fig. 2)

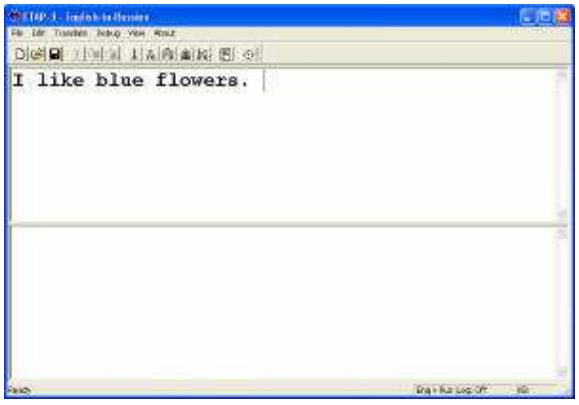


Fig. 2

When it comes to translating the word *blue,* the system finds that there are two options and no way to choose among them and activates the dialogue (Fig. 3).

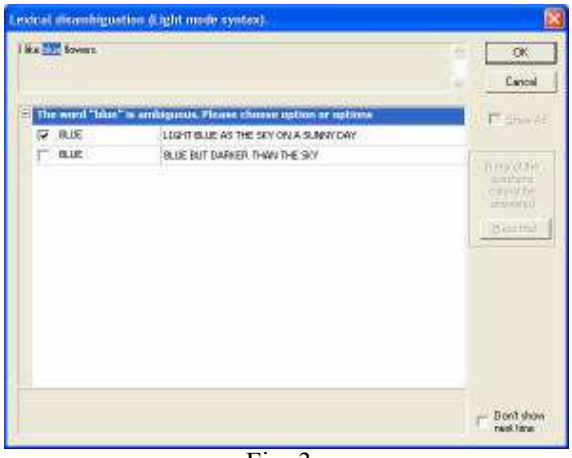


Fig. 3

In the dialogue box each option is provided with a short comment and/or example that helps the user choose among them. The user has to click the appropriate option (in Fig. 3 'light blue' is selected) and the system moves on. The result of the translation of this sentence is shown in Fig. 4.

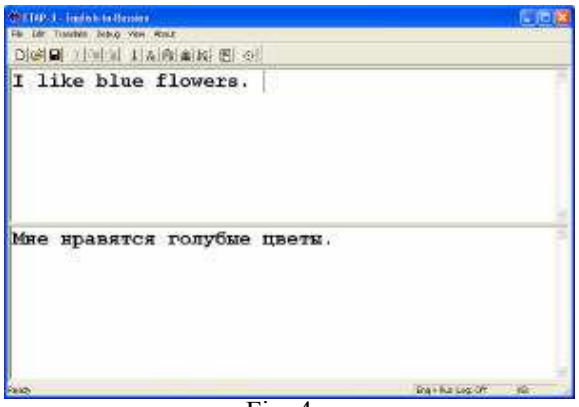


Fig. 4

Should we have selected the other option in the dialogue in Fig. 3, the result would have been different (Fig. 5).

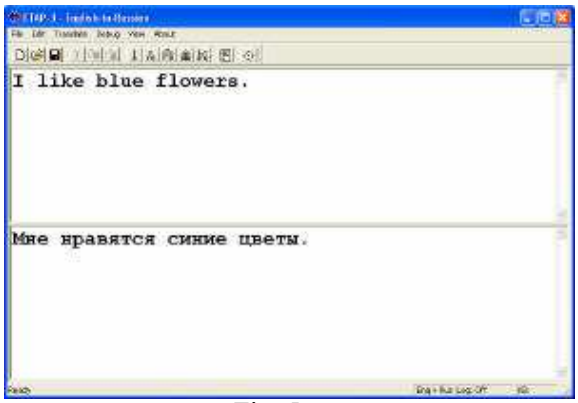


Fig. 5

It is important to note that the interactive disambiguation mode fully corresponds to the multiple translation possibilities discussed in the previous section. In particular, the dialogue takes into account all types of alternative translations irrespective of the way they are presented in the dictionary. It can be lexical or syntactic ambiguity that manifests itself in different lexico-syntactic structures of the source sentence, one-word translation variants within the same lexical meaning (of the *adjuration* type discussed above) or more complex phrases that translate a source word (of the *adventurer* type above).

3 UNL module in ETAP

One of ETAP-3 options is translation between Russian and the Universal Networking Language (UNL), put forward by H. Uchida of the United Nations University. Full specification of UNL and references to publications can be found at *http://www.undl.org.*

UNL is a formal language intended to represent information in a way that allows the generation of a text expressing this information in a large number of natural languages. A UNL expression is an oriented hyper-graph that corresponds to a NL sentence in the amount of information conveyed. The arcs are interpreted as semantic relations like *agent, object, time, place, manner,* etc. The nodes are special units, the so-called Universal Words (UW), interpreted as concepts, or groups of UWs. The concepts are built on the basis of English. When needed, English concepts can be modified by means of semantic restrictions in order to match better with the concepts of other languages. The nodes can be supplied with attributes which provide additional information on their use in the given sentence, e.g. *@imperative, @generic, @future, @obligation.*

3.1 Architecture

Since ETAP-3 is an NLP system based on rich linguistic knowledge, it is natural to maximally re-use its knowledge base and the whole architecture of the system in all applications. Our approach to UNL (described in Boguslavsky et al. 2000) is to build a bridge between UNL and one of the internal representations of ETAP, namely Normalized Syntactic Structure (NormSS), and in this way link UNL with all other levels of text representation, including the conventional orthographic form of the text.

The level of NormSS is best suited for establishing correspondence with UNL, as UNL expressions and NormSS show strong similarities. The most important of them are as follows:

a) Both UNL expressions and NormSSs occupy an intermediate position between the surface and the semantic levels of representation. They roughly correspond to the so-called deep-syntactic level. At this level the meaning of lexical items is not decomposed into semantic primitives, and the relations between lexical items are language independent.

b) The nodes of both UNL expressions and NormSSs are terminal elements (UWs in UNL vs. lexical items in NormSS) and not syntactic categories.

c) The nodes carry additional characteristics used in particular to convey grammatical information (attributes).

d) The arcs of both structures are non-symmetrical dependencies.

At the same time, UNL expressions and NormSSs differ in several important respects:

a) All nodes of NormSSs are lexical items, while a node of a UNL expression can be a sub-graph.

b) Nodes of a NormSS always correspond to one word sense, while UWs may either be broader or narrower than the corresponding English words.

c) A NormSS is a tree, while a UNL expression is a hyper-graph, which is a much more complicated object. Its arcs may form loops and connect sub-graphs.

d) The relations between the nodes in a NormSS are purely syntactic and are not supposed to convey a meaning of their own, while UNL relations denote semantic roles.

e) Attributes of a NormSS mostly correspond to grammatical elements, while UNL attributes often convey a meaning that is expressed in English or other natural languages by means of lexical items (e.g. modals).

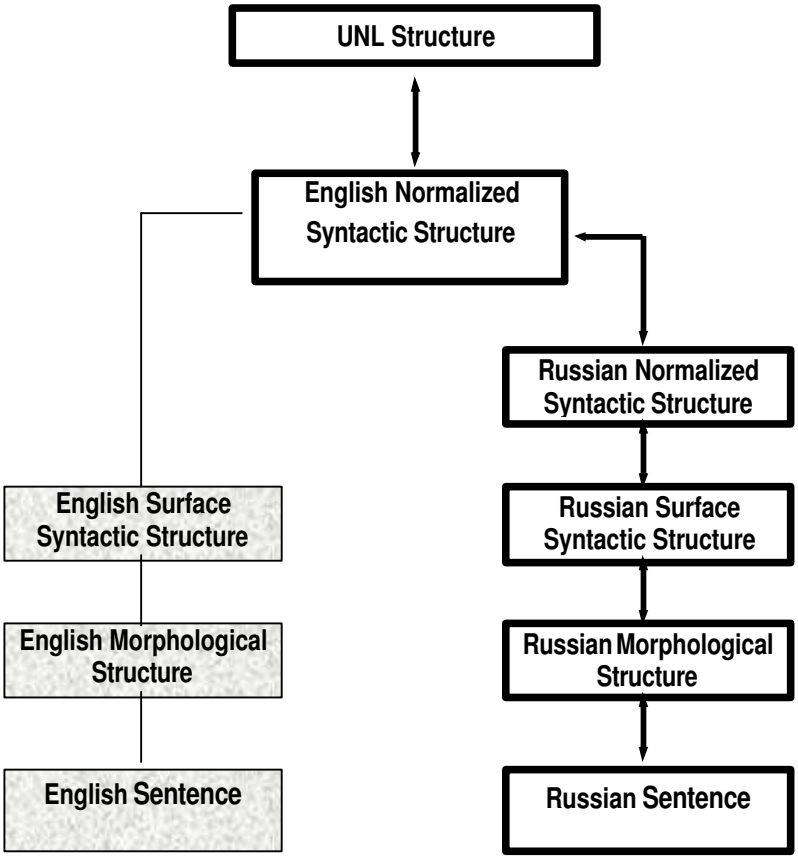


Fig. 6

f) A NormSS contains information on the word order, while a UNL expression does not say anything to this effect.

These differences and similarities make the task of establishing a bridge between UNL and NormSS far from trivial but feasible. Between the two types of NormSS readily available in ETAP – the Russian and the English one – we have chosen the latter, since it is the English concepts that serve for UNL as building blocks.

The architecture of the UNL module of ETAP-3 is given in Fig. 6.

3.2 UNL vs. English vs. Russian

As shown in Fig. 6, the interface between UNL and Russian is established at the level of the English NormSS. It ensures the maximum reuse of ETAP's English-to-Russian machine translation facility.

In the simple case, this scenario suggests that the UNL – Natural Language link can be localized within the English dictionary. This dictionary will only provide an English correspondence to UNL, which in most cases is not very difficult, and all the rest will be taken care of by the translation engine of ETAP. In this case, direct link between Russian and UNL is not needed at all, as long as ETAP covers the English-to-Russian correspondence.

However, the situation is not that simple. If we try to look at one language (Russian) through the perspective of another one (English), we encounter well-known problems. Let us illustrate the issue with an example. In Russian, there is no neutral equivalent of the English non-causative verb *to marry* as represented in sentences like *John married Ann in June.* The expression that exactly corresponds to this English verb – *vstupat' v brak* ('to contract a marriage') – is an official term and is not used in everyday life. Instead, Russian speakers make use of two different expressions: *zhenit'sja*, if the agent of the action is a male, and *vyxodit' zamuzh*, if it is a female. Since the English and the Russian words differ in their meaning, they correspond to different UWs. The UW for English *to marry* looks like (1), while Russian expressions have UNL equivalents with a more narrow meaning – (2) and (3), respectively (for simplicity's sake, only the relevant fragments of the UWs are given):

(1) marry(agt>human)
(2) marry(agt>male)
(3) marry(agt>female)

(Here agt stands for "agent").

Suppose the UNL expression that we receive at the input of our generator contains UW (2). Since we have to pass through English, we must first translate this concept into English and then translate the English word into Russian. But English has no direct equivalent of (2). It only has a word with a more general meaning – *to marry.* If our objective were to get the English text, this word would be perfectly in place. But since our target language is Russian, we cannot stop here and have to make a difficult choice between two different Russian equivalents.

This is exactly the problem that faces any translator from English into Russian, human or machine. Sometimes such a problem can be easily solved with the help of the context, sometimes it is less easy to solve or even unsolvable. For example, in the case of *blue* vs. *goluboj – sinij* discussed in 2.4 the context would hardly help to choose an appropriate Russian translation. However, in our example (2) the UNL source expression provides unambiguous information that allows avoiding this problem altogether, since the UW has only one correlate in Russian. If we pass from UNL to English and lose sight of the UNL source, we will lose the control of the semantic information and the quality of the output will deteriorate. This should not be permitted. Our solution to this problem is presented in 3.3.

In view of the above, it may seem that a better idea would be to sacrifice the benefit of reuse and establish a direct link between UNL and Russian.

However, the architecture shown in Fig. 6 has two more advantages that seem crucial.

First, this architecture allows us to make the UNL module of ETAP multilingual, that is to link UNL not only with Russian but also with English. In view of this perspective, it is reasonable to produce a full-fledged English NormSS that is much closer to UNL than the Russian one.

Second, the stock of the UNL concepts is continuously growing through the contributions coming from diverse languages. The UNL dictionaries of different languages grow at different rates and in different directions. Very often, the generator of language L_1 receives the UNL input produced by the UNL group of language L_2 that contains UWs that are absent from the UNL-to-L_1 dictionary. This happens particularly often with the so called multi-word UWs of the type

(4) International Research and Training Institute for the Advancement of Women (pof>General Assembly {(pof>United Nations)}).

If our only source of lexical knowledge were the UNL – Russian dictionary, we would not be

able to interpret such UWs, had they not been introduced in this dictionary in advance.

Our UNL-to-English architecture provides a universal solution to all difficulties of this kind. If the UW is not listed in the UNL dictionariy of ETAP, it is analyzed by means of the ETAP English dictionary and, if it is a multi-word expression, the English parser, which results in a reasonably good representation of the UW.

Moreover, it is often possible to correctly translate a UW that is absent from ETAP's UNL dictionary even if its headword is ambiguous. For example, if we receive UW

(5) open(mod<thing)

and do not find it in our UNL dictionary, we can replace it with the English word that stands in the position of the headword, that is *open*. However, this headword is ambiguous. In ETAP's English dictionary there are three entries for *open* - the adjective, the verb and the noun. A simple rule allows selecting the correct entry on the basis of the UW restriction: (mod<thing) means that the headword serves as a modifier of things. Hence, its English correlate is an adjective and not a verb or a noun.

3.3 UNL dictionary vs. English dictionary vs. Russian dictionary

The UNL-related information is distributed among the three ETAP dictionaries: UNL, English and Russian. The general idea is to combine (a) the idea of having the English NormSS as an intermediate level between UNL and the Russian NormSS and as a source of Russian and English generation and (b) the requirement of adequately treating cases of non-isomorphism between the English and the Russian concepts.

As shown in section 2.1, the ETAP dictionary entry contains several bilingual sub-zones, according to the number of working languages. In particular, the Russian dictionary has sub-zones for English and UNL, the English dictionary – for Russian and UNL and the UNL dictionary – for English and Russian.

Let us consider two cases: (1) the Russian and the English words are synonymous (as, for example, *to divorce* and *razvodit'sja*) and (2) they are not synonymous (as, for example, *to marry* and *zhenit'sja*).

The relevant fragments of the dictionary entries (with some simplifications) are as follows.

UNL dictionary:

NAME: divorce(agt>human)
ZONE:EN
TRANS: divorce
ZONE:RU
<none>

NAME: marry(agt>human)
ZONE:EN
TRANS: marry
ZONE:RU
<none>
NAME: marry(agt>male)
ZONE:EN
<none>
ZONE:RU
TRANS: zhenit'sja

English dictionary

NAME: divorce
ZONE: RU
TRANS: razvodit'sja
ZONE:UNL
TRANS: divorce(agt>human)
NAME: marry
ZONE: RU
TRANS: zhenit'sja / vyxodit' zamuzh
ZONE:UNL
TRANS: marry(agt>human)

Russian dictionary

NAME: razvodit'sja
ZONE: EN
TRANS: divorce
ZONE:UNL
TRANS: divorce(agt>human)
NAME: zhenit'sja
ZONE: EN
TRANS: marry
ZONE:UNL
TRANS: marry(agt>human)

Suppose we have to process a UNL expression that contains UW "divorce(agt>human)". Since this concept corresponds to both English and Russian words, we can do safely without any information on the Russian word in the UNL dictionary and obtain the NormSS with English *to divorce* taken from the English zone of the UNL entry. This NormSS allows generating both English and Russian texts by means of the standard ETAP transfer and generation facilities.

Let us consider the source UNL expression that contains UW "marry(agt>human)". It may have come from the language that, like English, German or Spanish, but unlike Russian or Polish, does not distinguish between the male-marriage and the female-marriage. The UNL dictionary entry for this UW will have the English translation but no Russian one, since Russian has no direct correlate for this concept. The problem of finding an appropriate Russian term is shifted to the level of the NormSS. At this level, we will have to find an equivalent of English *to marry*, just as if we translated from English and not from UNL. In this

case, the UNL source does not help us make a choice between two types of marriage. What does help is the mechanism of the interactive resolution of translational ambiguity described above, in 2.4.

Finally, let us examine the most interesting case - a UNL expression with UW "marry(agt>male)". The dictionary entry of this UW is symmetric to the entry of "marry(agt>human)": it contains a Russian correlate but no English one. In this situation, both English and Russian generations are not quite straightforward. As there is no direct English equivalent of this UW, the translation should be found by means of the UNL Knowledge Base (Uchida, 2003). In the absence of the operational version of KB, the general solution for processing an unknown UW is to extract the headword of the UW (*marry*) and treat it as an English word (cf. above, 3.2). This solves the problem of the generation of the English text. As for Russian, *zhenit'sja* indicated in the Russian zone of the UW entry is attached as a feature to the English node *marry*. At the stage of transfer from NormSS-English to NormSS-Russian, this feature will be lexicalized and replace the word *marry*.

4 Conclusion

The organization of lexical resources of the ETAP system allows reusing the dictionaries in diverse applications, such as machine translation in various language pairs and translation to and from UNL. In all the applications, there are three modes of operation supported by the dictionaries: automatic production of a single (most probable) translation, automatic production of all possible translations and the interactive translation with the dialogue-based disambiguation.

References

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Mitjushin L.G., Sannikov, V.Z., Cinman, L.L. (1992) *Lingvisticheskij processor dlja slozhnyx informacionnyx sistem. [A linguistic processor for advanced information systems.]* Moskva, Nauka. 256 p.

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Sannikov V.Z. and Tsinman L.L. 1992b. The Linguistics of a Machine Translation System. *Meta*, 37 (1): 97-112.

Apresjan Ju.D., Boguslavskij I.M., Iomdin L.L., Lazurskij A.V., Sannikov V.Z. and Tsinman L.L. 1993. *Systeme de traduction automatique {ETAP}*. In: La Traductique. P.Bouillon and A.Clas (eds). Montreal, Les Presses de l'Universite de Montreal.

Apresjan, Ju.D. 1995. *Integral'noe opisanie jazyka i sistemnaja leksikografija [An Integrated Description of Language and Systematic lexicography.]* Moscow, Jazyki russkoj kul'tury.

Apresjan, Ju. D. 2000. *Systematic Lexicography.* Oxford University Press, London, 304 p.

Apresian Ju., I. Boguslavsky, L. Iomdin, A. Lazursky, V. Sannikov, V. Sizov, L. Tsinman. 2003. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. In: *MTT 2003, First International Conference on Meaning – Text Theory*. Paris, Ecole Normale Superieure, Paris, 279-288.

Blanchon, H. Interagir pour traduire: la TAO personnelle pour redacteur monolingue. *La Tribune des Industries de la Langues. Vol. 17-18-19, 1995, pp. 28-34.*

Blanchon, H. A Customizable Interactive Disambiguation Methodology and Two Implementations to Disambiguate French and English Input. *Proc. MIDDIM'96.* Le col de porte, Isere, France. 12-14 Aout 1996. Vol. 1/1, 1996, pp. 190-200.

Blanchon, H. Interactive Disambguation of Natural Language Input: a Methodology and Two Implementations for French and English. *Proc. IJCAI-97.* Nagoya, Japan. August 23-29, 1997. Vol. 2/2, 1997, pp. 1042-1047

Boguslavsky I., N. Frid, L. Iomdin, L. Kreidlin, I. Sagalova, V. Sizov. 2000. Creating a Universal Networking Language Module within an Advanced NLP System. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, 2000, 83-89.

Boguslavsky I., L. Iomdin, V. Sizov. 2003. Interactive enconversion by means of the ETAP-3 system. In "Proceedings of the International Conference on the Convergence of Knowledge, Culture, Language and Information Technologies", Alexandria, 2003.

Boitet, C. & Blanchon, H. Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup. *Machine Translation. Vol. 9(2), 1995, pp 99-132.*

Hutchins W. J., H. L. Somers. 1992. *An Introduction to Machine Translation.* Academic Press, London.

Mel'cuk I. 1974. *Opyt teorii lingvisticheskix modelej "Smysl – Tekst".* Moscow, "Nauka" Publishers.

Uchida H. 2003. *The UW Manual.* http://www.undl.org.

A Model for Fine-Grained Alignment of Multilingual Texts

Lea CYRUS and Hendrik FEDDES*
Arbeitsbereich Linguistik
University of Münster
Hüfferstraße 27, 48149 Münster, Germany
{lea,feddes}@marley.uni-muenster.de

Abstract

While alignment of texts on the sentential level is often seen as being too coarse, and word alignment as being too fine-grained, bi- or multilingual texts which are aligned on a level in-between are a useful resource for many purposes. Starting from a number of examples of non-literal translations, which tend to make alignment difficult, we describe an alignment model which copes with these cases by explicitly coding them. The model is based on predicate-argument structures and thus covers the middle ground between sentence and word alignment. The model is currently used in a recently initiated project of a parallel English-German treebank (FuSe), which can in principle be extended with additional languages.

1 Introduction

When building parallel linguistic resources, one of the most obvious problems that need be solved is that of alignment. Usually, in sentence- or word-aligned corpora, alignments are unmarked relations between corresponding elements. They are unmarked because the kind of correspondence between two elements is either obvious or beyond classification. E. g., in a sentence-aligned corpus, the $n : m$ relations that hold between sentences express the fact that the propositions contained in n sentences in L1 are basically the same as the propositions in m sentences in L2 (lowest common denominator). No further information about the kind of correspondence could possibly be added on this degree of granularity. On the other hand, in word-aligned corpora, words are usually aligned as being "lexically equivalent" or are not aligned at all.[1] Although there are many shades of "lexical equivalence", these are usually not explicitly categorised. As (Hansen-Schirra and Neumann, 2003) point out, for many research questions neither type of alignment is sufficient, since the most interesting phenomena can be found on a level between these two extremes.

We propose a more finely grained model of alignment which is based on monolingual predicate-argument structures, since we assume that, while translations can be non-literal in a variety of ways, they must be based on similar predicates and arguments for some kind of translational equivalence to be achieved. Furthermore, our model explicitly encodes the ways in which the two versions of a text deviate from each other. (Salkie, 2002) points out that the possibility to investigate what types of non-literal translations occur on a regular basis is one of the major profits that linguists and translation theorists can draw from parallel corpora.

In Section 2, we begin by describing some ways in which translations can deviate from one another. We then describe in detail the alignment model, which is based on a monolingual predicate-argument structure (Section 3). In Section 4 we conclude by introducing the parallel treebank project FuSe which uses the model described in this paper to align German and English texts from the Europarl parallel corpus (Koehn, 2002).

2 Differences in Translations

In most cases, translations are not absolutely literal counterparts of their source texts. In order to avoid translationese, i. e. deviations from the norms of the target language, a skilled translator will apply certain mechanisms, which (Salkie, 2002) calls "inventive translations" and which need to be captured and systematised. The following section will give some examples[2]

* We would like to thank our colleague Frank Schumacher for many valuable comments on this paper.

[1] Cf. the approach described in (Melamed, 1998).

[2] As we work with English and German, all examples are taken from these two languages. They are taken from the Europarl corpus (see Section 4) and are abbreviated where necessary. Unfortunately, it is not eas-

of common discrepancies encountered between a source text and its translation.

2.1 Nominalisations

Quite frequently, verbal expressions in L1 are expressed by corresponding nominalisations in L2. This departure from the source text results in a completely different structure of the target sentence, as can be seen in (1) and (2), where the English verb *harmonise* is expressed as *Harmonisierung* in German. The argument of the English verb functioning as the grammatical subject is realised as a postnominal modifier in the German sentence.

(1) The laws against racism must be harmonised.[3]

(2) Die Harmonisierung der
The harmonisation of_the
Rechtsvorschriften gegen den
laws against the
Rassismus ist dringend erforderlich.
racism is urgently necessary.

This case is particularly interesting, because it involves a case of modality. In the English sentence, the verb is modified by the modal auxiliary *must*. In order to express the modality in the German version, a different strategy is applied, namely the use of an adjective with modal meaning (*erforderlich*, 'necessary'). Consequently, there are two predications in the German sentence as opposed to only one predication in the English sentence.

2.2 Voice

A further way in which translations can differ from their source is the choice of active or passive voice. This is exemplified by (3) and (4). Here, the direct object of the English sentence corresponds to the grammatical subject of the German sentence, while the subject of the English sentence is realised as a prepositional phrase with *durch* in the German version.

(3) The conclusions of the Theato report safeguard them perfectly.[4]

(4) Durch die Schlußfolgerungen des
By the conclusions of_the
Berichts Theato werden sie
report Theato are they
uneingeschränkt bewahrt.
unlimitedly safeguarded

2.3 Negation

Sometimes, a positive predicate expression is translated by negating its antonym. This is the case in (5) and (6): both sentences contain a negative statement, but while the negation is incorporated into the English adjective by means of the negative prefix *in-*, it is achieved syntactically in the German sentence.

(5) the Directive is inapplicable in Denmark[5]

(6) die Richtlinie ist in Dänemark nicht
the Directive is in Denmark not
anwendbar
applicable

2.4 Information Structure

Sentences and their translations can be organised differently with regard to their information structure. Sentences (7) and (8) are a good example for this type of non-literal translation.

(7) Our motion will give you a great deal of food for thought, Commissioner[6]

(8) Eine Reihe von Anregungen werden
A row of suggestions will
wir Ihnen, Herr Kommissar, mit
we you, Mr. Commissioner, with
unserer Entschließung mitgeben
our resolution give

The German sentence is rather inconspicuous, with the grammatical subject being a prototypical agent (*wir*, 'we'). In the English version, however, it is the means that is realised in subject position and thus perspectivised. The corresponding constituent in German (*mit unserer Entschließung*, 'with our motion') is but an adverbial. In English, the actual agent is not realised as such and can only be identified by a process of inference based on the presence of the possessive pronoun *our*. Thus, while being more or less equivalent in meaning, this sentence pair differs significantly in its overall organisation.

ily discernible from the corpus data which language is the source language. Consequently, our use of the terms 'source', 'target', 'L1', and 'L2' does not admit of any conclusions as to whether one of the languages is the source language, and if so, which one.

[3]Europarl:de-en/ep-00-01-19.al, 489.

[4]Europarl:de-en/ep-00-01-18.al, 749.

[5]Europarl:de-en/ep-00-01-18.al, 2522.

[6]Europarl:de-en/ep-00-01-18.al, 53.

3 Alignment Model

The alignment model we propose is based on the assumption that a representation of translational equivalence can best be approximated by aligning the elements of monolingual predicate-argument structures. Section 3.1 describes this layer of the model in detail and shows how some of the differences in translations described in Section 2 can be accomodated on such a level. We assume that the annotation model described here is an extension to linguistic data which are already annotated with phrase-structure trees, i. e. treebanks. Section 3.2 shows how the binding of predicates and arguments to syntactic nodes is modelled. Section 3.3 describes the details of the alignment layer and the tags used to mark particular kinds of alignments, thus accounting for some more of the differences shown in Section 2.

3.1 Predicates and Arguments

The predicate-argument structures used in our model consist solely of predicates and their arguments. Although there is usually more than one predicate in a sentence, no attempt is made to nest structures or to join the predications logically in any way. The idea is to make the predicate-argument structure as rich as is necessary to be able to align a sentence pair while keeping it as simple as possible so as not to make it too difficult to annotate. In the same vein, quantification, negation, and other operators are not annotated. In short, the predicate-argument structures are not supposed to capture the semantics of a sentence exhaustively in an interlingua-like fashion.

To have clear-cut criteria for annotators to determine what a predicate is, we rely on the heuristic assumption that predicates are more likely to be expressed by tokens belonging to some word classes than by tokens belonging to others. Potential predicate expressions in this model are verbs, deverbal adjectives and nouns[7] or other adjectives and nouns which show a syntactic subcategorisation pattern. The predicates are represented by the capitalised citation form of the lexical item (e. g. HARMONISE). They are assigned a class based on their syntactic form (*v*, *n*, *a* for 'verbal', 'nominal', and 'adjectival', respectively), and derivationally related predicates form a predicate group.

Arguments are given short intuitive role names (e. g. ENT_HARMONISED, i. e. the entity being harmonised) in order to facilitate the annotation process. These role names have to be used consistently only within a predicate group. If, for example, an argument of the predicate HARMONISE has been assigned the role ENT_HARMONISED and the annotator encounters a comparable role as argument to the predicate HARMONISATION, the same role name for this argument has to be used.[8]

The usefulness of such a structure can be shown by analysing the sentence pair (1) and (2) in Section 2.1. While the syntactic constructions differ considerably, the predicate-argument structure shows the correspondence quite clearly (see the annotated sentences in Figure 1[9]): in the English sentence, we find the predicate HARMONISE with its argument ENT_HARMONISED, which corresponds to the predicate HARMONISIERUNG and its argument HARMONISIERTES in the German sentence. The information that a predicate of the class *v* is aligned with a predicate of the class *n* can be used to query the corpus for this type of non-literal translations.

The active vs. passive translation in sentences (3) and (4) is another phenomenon which is accomodated by a predicate-argument structure (Figure 2): the subject NP_{502} in the English sentence corresponds to the passivised subject NP_{502} (embedded in PP_{503}) in the German sentence on the basis of having the same argument role (SAFEGUARDER vs. BEWAHRER) in a comparable predication.

It is sometimes assumed that predicate-argument structure can be derived or recovered from constituent structure or functional tags such as subject and object.[10] It is true that these annotation layers provide important heuristic clues for the identification of predi-

[7] For all non-verbal predicate expressions for which a derivationally related verbal expression exists it is assumed that they are deverbal derivations, etymological counter-evidence notwithstanding.

[8] Keeping the argument names consistent for all predicates within a group while differentiating the predicates on the basis of syntactic form are complementary principles, both of which are supposed to facilitate querying the corpus. The consistency of argument names within a group, for example, enables the researcher to analyse paradigmatically all realisations of an argument irrespective of the syntactic form of the predicate. At the same time, the differentiation of predicates makes possible a syntagmatic analysis of the differences of argument structures depending on the syntactic form of the predicate.

[9] All figures are at the end of the paper.

[10] See e. g. (Marcus et al., 1994).

cates and arguments and may eventually speed up the annotation process in a semi-automatic way. But, as the examples above have shown, predicate-argument structure goes beyond the assignment of phrasal categories and grammatical functions, because the grammatical category of predicate expressions and consequently the grammatical functions of their arguments can vary considerably. Also, the predicate-argument structure licenses the alignment relation by showing explicitly what it is based on.

3.2 Binding Layer

As mentioned above, we assume that the annotation model described here is used on top of syntactically annotated data. Consequently, all elements of the predicate-argument structure must be bound to elements of the phrasal structure (terminal or non-terminal nodes). These bindings are stored in a dedicated binding layer between the constituent layer and the predicate-argument layer.

A problem arises when there is no direct correspondence between argument roles and constituents. For instance, this is the case whenever a noun is postmodified by a participle clause: in Figure 3, the argument role ENT_RAISED of the predicate RAISE is realised by NP_{525}, but the participle clause (IPA_{517}) containing the predicate ($raised_6$) needs to be excluded, because not excluding it would lead to recursion. Consequently, there is no simple way to link the argument role to its realisation in the tree.

In these cases, the argument role is linked to the appropriate phrase (here: NP_{525}) and the constituent that contains the predicate (IPA_{517}) is pruned out, which results in a discontinuous argument realisation. Thus, in general, the binding layer allows for complex bindings, with more than one node of the constituent structure to be included in and sub-nodes to be explicitly excluded from a binding to a predicate or argument.[11]

When an expected argument is absent on the phrasal level due to specific syntactic constructions, the binding of the predicate is tagged accordingly, thus accounting for the missing argument. For example, in passive constructions like in Table 1, the predicate binding is tagged as `pv`. Other common examples are imperative constructions. Although information of this kind may possibly be derived from the constituent structure, it is explicitly recorded in the binding layer as it has a direct impact on the predicate-argument structure and thus might prove useful for the automatic extraction of valency patterns.

Sentence	wenn	korrekt	gedolmetscht	wurde
Gloss	if	correctly	interpreted	was
			↑	
Binding			pv	
			|	
Pred/Arg			DOLMETSCHEN	

Table 1: Example of a tagged predicate binding (Europarl:de-en/ep-00-01-18.al, 2532)

Note that the passive tag can also be exploited in order to query for sentence pairs like (3) and (4) (in Section 2.2), where an active sentence is translated with a passive: it is straightforward to find those instances of aligned predicates where only one binding carries the passive tag.

3.3 Alignment Layer

On the alignment layer, the elements of a pair of predicate-argument structures are aligned with each other. Arguments are aligned on the basis of corresponding roles within the predications. Comparable to the tags used in the binding layer that account for specific constructions (see Section 3.2), the alignments may also be tagged with further information. These tags are used to classify types of non-literalness like those discussed in Sections 2.3 and 2.4.[12]

Sentences (5) and (6) are an example for a tagged alignment. As Section 2.3 has shown, negation may be incorporated in a predicate in L1, but not in L2. Since our predicate-argument structure does not include syntactic negation, this results in the alignment of a predicate in L1 with its logical opposite in L2. To account for this fact, predicate alignments of this kind are tagged as absolute opposites (`abs-opp`).

Similarly, alignment tagging is applied when predications are in some way incompatible, as is the case with sentences (7) and (8) in Section 2.4. As can be seen in the aligned annotation (Figure 4), the different information structure of these sentences has caused the two corresponding argument roles of GIVER and MITGEBER to be realised by two incompatible expressions representing different referents (NP_{500}

[11] See the database documentation (Feddes, 2004) for a more detailed description of this mechanism.

[12] The deviant translations described in Sections 2.1 and 2.2 are already represented via predicate class (see Section 3.1) and on the binding layer (see Section 3.2), respectively.

vs. wir_5). In this case, the alignment between the incompatible arguments is tagged `incomp`.

If there is no corresponding predicate-argument structure in the other language (as e. g. the adjectival predicate in sentence (2)) or if an argument within a structure does not have a counterpart in the other language, there will be no alignment.

Table 2 gives an overview of the annotation layers as described in this section.

Layer	*Function*
Phrasal	constituent structure of language A
Binding	binding ↓ predicates/arguments to ↑ nodes
PA	predicate-argument structures
Alignment	aligning ↕ predicates and arguments
PA	predicate-argument structures
Binding	binding ↑ predicates/arguments to ↓ nodes
Phrasal	constituent structure of language B

Table 2: The layers of the predicate-argument annotation

All elements of the alignment structure are supposed to mark explicitly the way they contribute to or distort the resulting translational equivalence of a sentence pair.[13] First and foremost, if two elements are aligned to each other, this alignment is licensed by their having comparable roles in the predicate-argument structures. This is the default case. If, however, a particular alignment relation, either of predicates or of arguments, is deviant in some way, this deviance is explicitly marked and classified on the alignment layer.

4 Application and Outlook

The alignment model we have described is currently being used in a project to build a treebank of aligned parallel texts in English and German with the following linguistic levels: POS tags, constituent structure and functional relations, plus the predicate-argument structure and the alignment layer to "fuse" the two – hence our working title for the treebank, FuSe, which additionally stands for *fu*nctional *se*mantic annotation (Cyrus et al., 2003; Cyrus et al., 2004).

Our data source, the Europarl corpus (Koehn, 2002), contains sentence-aligned proceedings of the European parliament in eleven languages and thus offers ample opportunity for extending the treebank at a later stage.[14] For syntactic and functional annotation we basically adapt the TIGER annotation scheme (Albert and others, 2003), making adjustments where we deem appropriate and changes which become necessary when adapting to English an annotation scheme which was originally developed for German.

We use ANNOTATE for the semi-automatic assignment of POS tags, hierarchical structure, phrasal and functional tags (Brants, 1999; Plaehn, 1998a). ANNOTATE stores all annotations in a relational database.[15] To stay consistent with this approach we have developed an extension to the ANNOTATE database structure to model the predicate-argument layer and the binding layer.

Due to the monolingual nature of the ANNOTATE database structure, the alignment layer (Section 3.3) cannot be incorporated into it. Hence, additional types of databases are needed. For each language pair (currently English and German), an alignment database is defined which represents the alignment layer, thus fusing two extended ANNOTATE databases. Additionally, an administrative database is needed to define sets of two ANNOTATE databases and one alignment database. The final parallel treebank will be represented by the union of these sets (Feddes, 2004).

While annotators use ANNOTATE to enter phrasal and functional structures comfortably, the predicate-argument structures and alignments are currently entered into a structured text file which is then imported into the database. A graphical annotation tool for these layers is under development. It will make binding the predicate-argument structure to the constituent structure easier for the annotators and suggest argument roles based on previous decisions.

Possiblities of semi-automatic methods to speed up the annotation and thus reduce the costs of building the treebank are currently being investigated.[16] Still, quite a bit of manual

[13] Cf. the "translation network" described in (Santos, 2000) for a much more complex approach to describing translation in a formal way; this model, however, goes well beyond what we think is feasible when annotating large amounts of data.

[14] There are a few drawbacks to Europarl, such as its limited register and the fact that it is not easily discernible which language is the source language. However, we believe that at this stage the easy accessibility, the amount of preprocessing and particularly the lack of copyright restrictions make up for these disadvantages.

[15] For details about the ANNOTATE database structure see (Plaehn, 1998b).

[16] One track we follow is to investigate if it is feasible to

work will remain. We believe, however, that the effort that goes into such a gold-standard parallel treebank is very much worthwhile since the treebank will eventually prove useful for a number of fields and can be exploited for numerous applications. To name but a few, translation studies and contrastive analyses will profit particularly from the explicit annotation of translational differences. NLP applications such as Machine Translation could, e. g., exploit the constituent structures of two languages which are mapped via the predicate-argument-structure. Also, from the disambiguated predicates and their argument structures, a multilingual valency dictionary could be derived.

References

Stefanie Albert et al. 2003. TIGER Annotationsschema. Technical report, Universität des Saarlandes, Universität Stuttgart, Universität Potsdam. Unpublished Draft – 24 July 2003.

Thorsten Brants. 1999. *Tagging and Parsing with Cascaded Markov Models: Automation of Corpus Annotation*, volume 6 of *Saarbrücken Dissertations in Computational Linguistics and Language Technology*. Saarland University, Saarbrücken.

Lea Cyrus, Hendrik Feddes, and Frank Schumacher. 2003. FuSe – a multi-layered parallel treebank. Poster presented at the Second Workshop on Treebanks and Linguistic Theories, 14–15 November 2003, Växjö, Sweden (TLT 2003). `http://fuse.uni-muenster.de/Publications/0311_tltPoster.pdf`.

Lea Cyrus, Hendrik Feddes, and Frank Schumacher. 2004. Annotating predicate-argument structure for a parallel treebank. In Charles J. Fillmore, Manfred Pinkal, Collin F. Baker, and Katrin Erk, editors, *Proc. LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora, Lisbon, May 30, 2004*, pages 39–46. `http://fuse.uni-muenster.de/Publications/0405_lrec.pdf`.

Hendrik Feddes. 2004. FuSe database structure. Technical report, Arbeitsbereich Linguistik, University of Münster. `http://fuse.uni-muenster.de/Publications/dbStruktur.pdf`.

Silvia Hansen-Schirra and Stella Neumann. 2003. The challenge of working with multilingual corpora. In Stella Neumann and Silvia Hansen-Schirra, editors, *Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives. Corpus Linguistics 2003, Lancaster*, pages 1–6.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished draft, `http://www.isi.edu/~koehn/publications/europarl/`.

Mitch Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proc. ARPA Human Language Technology Workshop*.

I. Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. Technical Report 98-07, IRCS, University of Pennsylvania. `http://citeseer.ist.psu.edu/melamed98manual.html`.

Oliver Plaehn. 1998a. ANNOTATE Bedienungsanleitung. Technical report, Universität des Saarlandes, FR 8.7, Saarbrücken. `http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate-manual.ps.gz`.

Oliver Plaehn. 1998b. ANNOTATE Datenbank-Dokumentation. Technical report, Universität des Saarlandes, FR 8.7, Saarbrücken. `http://www.coli.uni-sb.de/sfb378/negra-corpus/datenbank.ps.gz`.

Raphael Salkie. 2002. How can linguists profit from parallel corpora? In Lars Borin, editor, *Parallel Corpora, Parallel Worlds*, pages 93–109. Rodopi, Amsterdam.

Diana Santos. 2000. The translation network: A model for a fine-grained description of translations. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*, volume 13 of *Text, Speech and Language Technology*, chapter 8. Kluwer, Dordrecht.

have the annotators mark predicate-argument structures on raw texts and have the phrasal and functional layers added in a later stage, possibly supported by methods which derive these layers partially from the predicate-argument structures. This is, however, still very tentative.

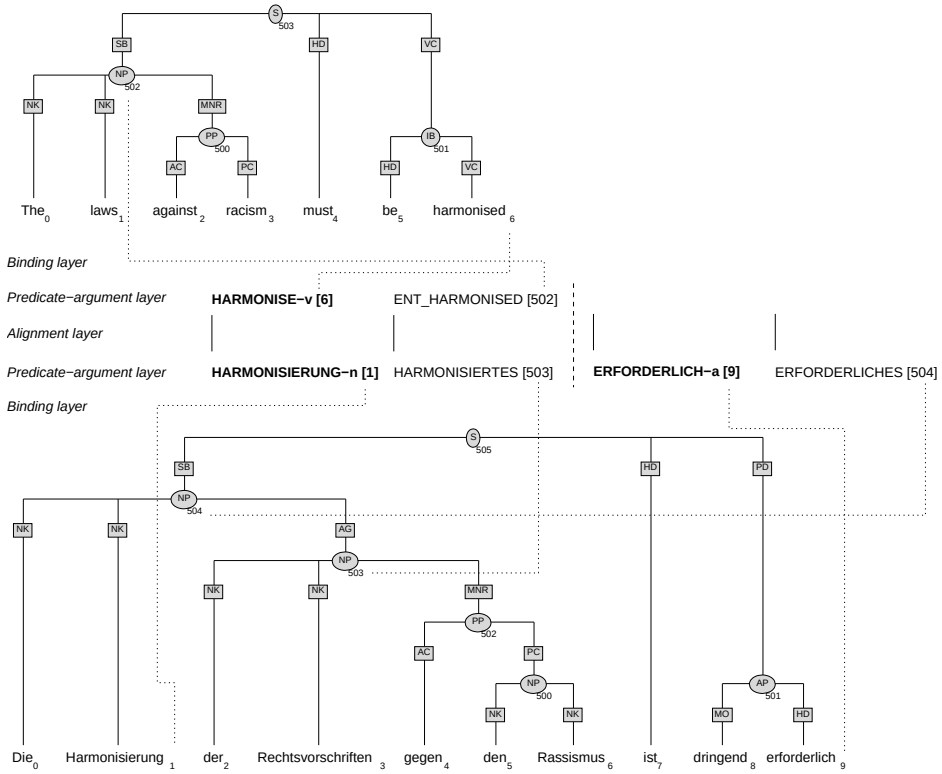


Figure 1: Alignment of a verb/direct-object construction with a noun/modifier construction

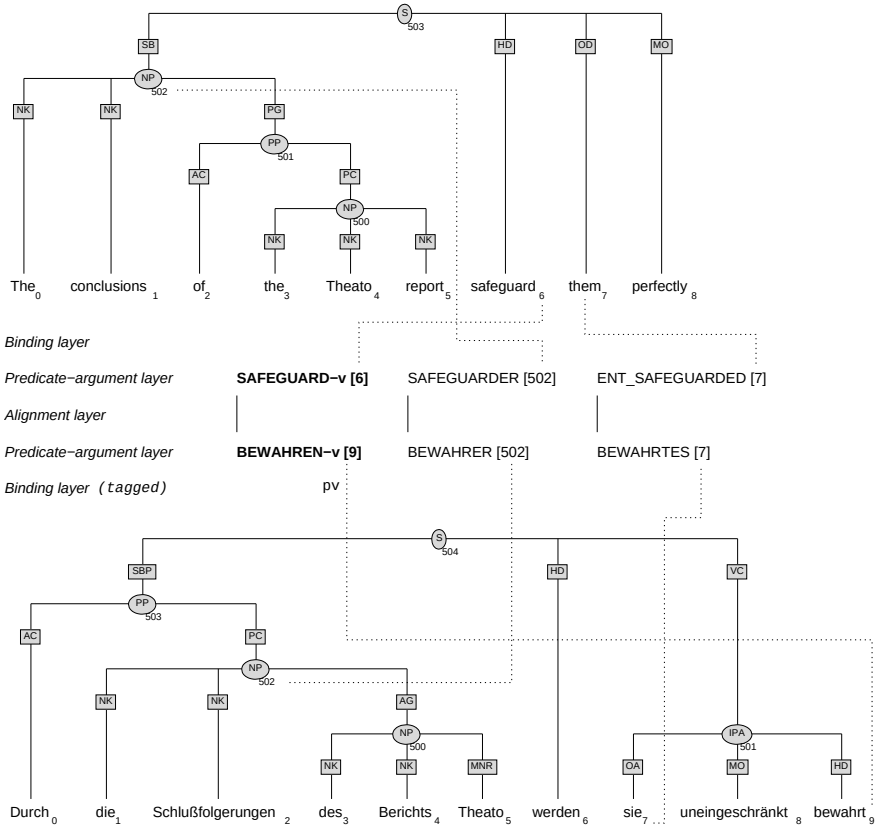


Figure 2: Active vs. passive voice in translations: an example of a tagged binding (`pv`)

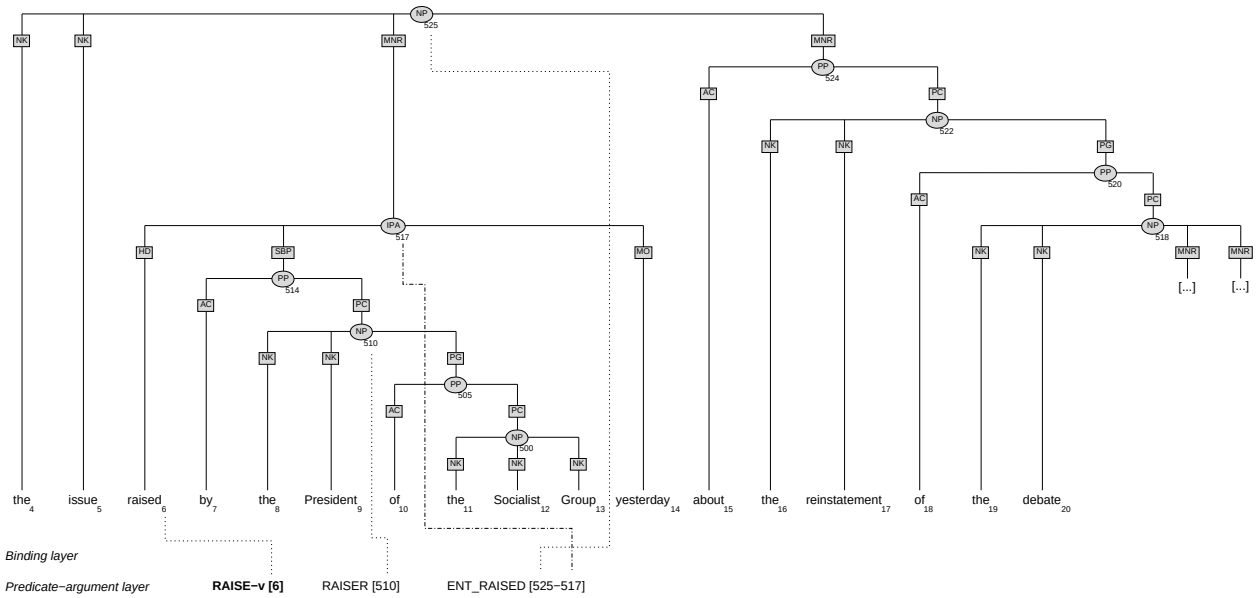


Figure 3: Complex binding of an argument: an example of a pruned constituent (dash-dotted line)

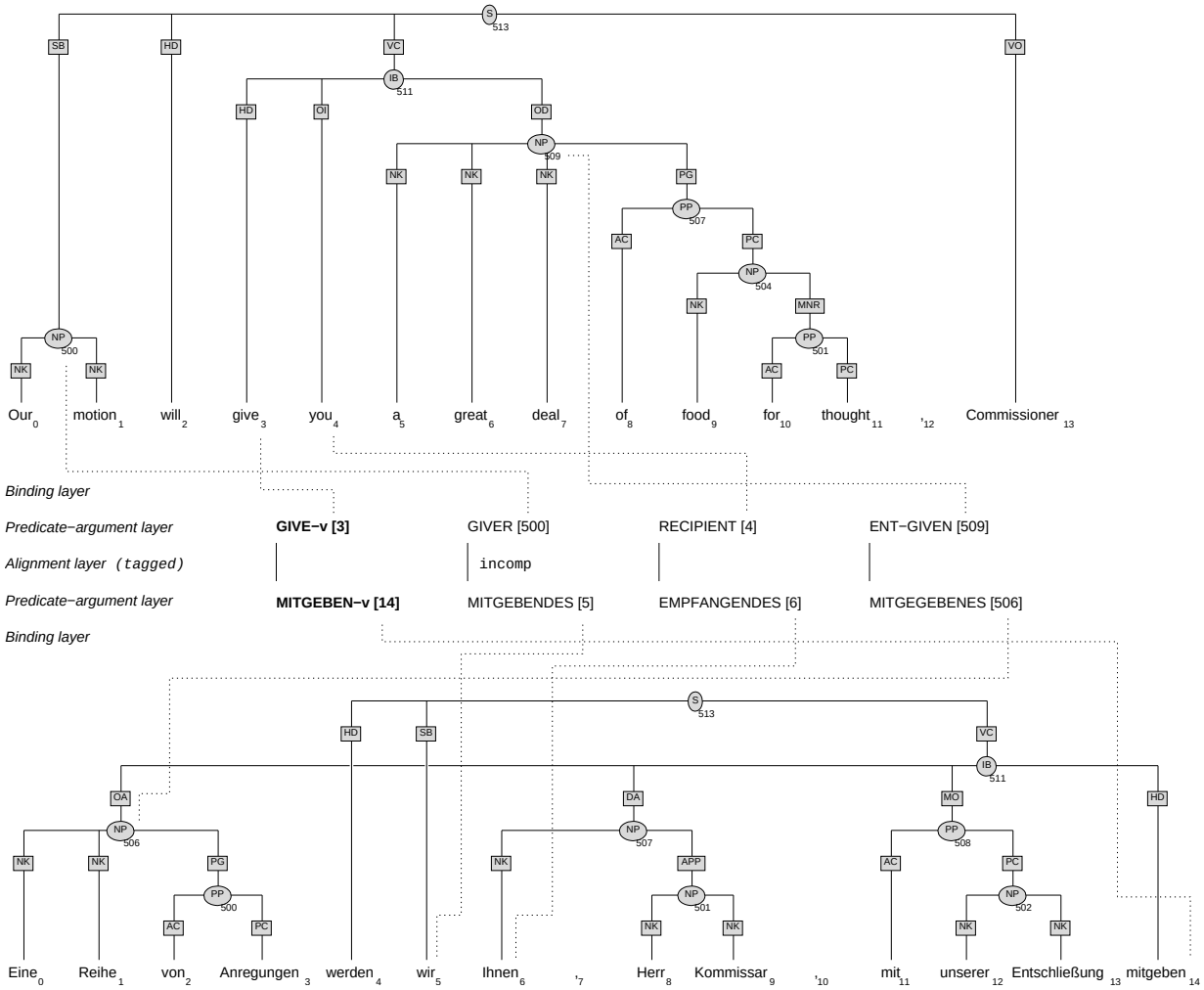


Figure 4: Different information structure: an example of a tagged alignment (`incomp`)

Qualitative Evaluation of Automatically Calculated Acception Based MLDB

Aree Teeraparbseree
GETA, CLIPS, IMAG
385, rue de la Bibliothèque
B.P. 53 - 38041 Grenoble Cedex 9, France
aree.teeraparbseree@imag.fr

Abstract

In the context of the Papillon project, which aims at creating a multilingual lexical database (MLDB), we have developed Jeminie, an adaptable system that helps automatically building interlingual lexical databases from existing lexical resources. In this article, we present a taxonomy of criteria for evaluating a MLDB, that motivates the need for arbitrary compositions of criteria to evaluate a whole MLDB. A quality measurement method is proposed, that is adaptable to different contexts and available lexical resources.

1 Introduction

The Papillon project[1] aims at creating a cooperative, free, permanent, web-oriented environment for the development and the consultation of a multilingual lexical database. The macrostructure of Papillon is a set of monolingual dictionaries (one for each language) of word senses, called *lexies*, linked through a central set of interlingual links, called *axies*. Axies, also called *interlingual acceptions*, are not concepts, but simply interlingual links between lexies, motived by translations found in existing dictionaries or proposed by the contributors. Figure 1 represents an interlingual database that links monolingual resources in three languages: French, English and Japanese. The interlingual acceptions (axies) are linked to lexies from each language. For instance, a lexie for the French word "terre" is linked through an axie to two lexies for the English words "earth" and "soil" and to a lexie for the Japanese word "tsuchi". Note that an axie can be refined into a set of axies. For instance, a lexie for the English word "chair" is linked through axie1 to two lexies for the French words "fauteuil" and "chaise". Axie1 can be refined into two axies axie11 and axie12 as illustrated in figure 2.

[1] http://www.papillon-dictionary.org/

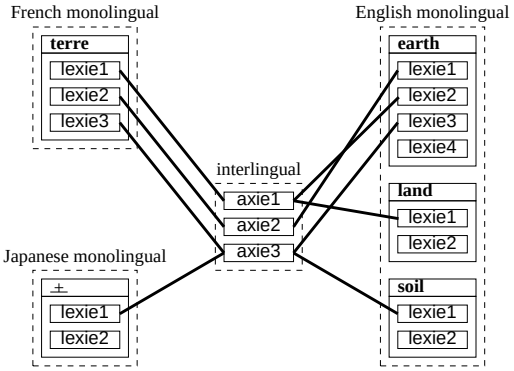


Figure 1: An example interlingual database

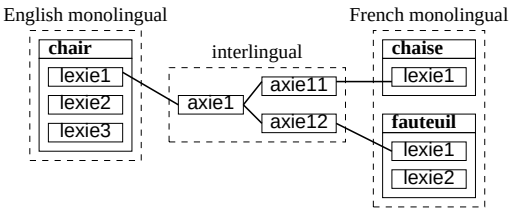


Figure 2: An example of refined axies

This pivot macrostructure has been defined by (Sérasset, 1994) and experimented by (Blanc, 1999) in the PARAX mockup. The microstructure of the monolingual dictionaries is the "DiCo" structure, which is a simplification of Mel'cuk's (Mel'cuk et al., 1995) DEC (Explanatory-Combinatorial Dictionary) designed by Polguère & Mel'cuk (Polguère, 2000) to make it possible to construct large, detailed and principled dictionaries in tractable time.

The building method of the Papillon lexical database is based on one hand on 1) reusing existing lexical resources, and on the other hand on 2) contributions of volunteers working through Internet. In order to automate the first step, we have developed Jeminie (cf. section 2), a

flexible software system that helps create (semi-) automatically interlingual lexical databases. As there are several possible techniques for the creation of axies that can be implemented in Jeminie, it is necessary to evaluate and compare these techniques to understand their strengths and weaknesses and to identify possible improvements. This article proposes an approach for the automatic qualitative evaluation of an automatically created MLDB, for instance created by Jeminie, that relies on an evaluation software system that adapts to the measured MLDB.

The next section of this article provides an overview of the Jeminie system and the strategy it implements to create interlingual lexical databases. The third section presents in detail evaluation criteria for an MLDB. The fourth section describes the evaluation system that we propose and the metrics and criteria to evaluate the quality of MLDB. Last sections discuss the measurement strategy and conclude.

2 Jeminie

Jeminie is a software system that helps building interlingual databases. Its first function is to automatically extract information from existing monolingual dictionaries, at least one for each considered language, and to *normalize* it into lexies. The second function of Jeminie is to automatically link lexies that have the same sense into axies. The prominent feature of Jeminie is the ability to arbitrarily *combine several axie creation techniques* (Teeraparbseree, 2003).

An axie creation technique is an *algorithm* that creates axies to link a set of existing lexies. An algorithm may use existing additional lexical resources, such as: bilingual dictionaries, parallel corpora, synonym dictionaries, and antonym dictionaries. Algorithms that do not rely on additional lexical resources consider only information available from the monolingual databases, and include vectorial algorithms such as calculating and comparing conceptual vectors for each lexie (Lafourcade, 2002).

The use of one algorithm alone is not sufficient, in practice, to produce a good quality MLDB. For instance, using only one algorithm that uses bilingual dictionaries, one obtains a lexical database on the level of words but not on the level of senses of words. The Jeminie system tackles this problem from a *software engineering* point of view. In Jeminie, an axie creation algorithm is implemented in a *reusable software module*. Jeminie allows for arbitrary composition of modules, in order to take advantage of each axie creation algorithm, and to create a MLDB of the best possible quality. We call a MLDB *production process*, a sequence of executions of axie creation modules. A process is specified using a specific language that provides high-level abstractions. The Jeminie architecture is divided into three layers. The *core layer* is a library that is used to implement axie creation modules at the *module layer*. The *processes interpreter* starts the execution of modules according to processes specified by linguists. The interpreter is developed using the core layer. Jeminie has been developed in Java following object-oriented design techniques and patterns.

Each execution of an axie creation module progressively contributes to create and filter the intermediate set of axies. The final MLDB is obtained after the last module execution in a process. The quality of a MLDB can be evaluated either 1) on the final set of axies after a whole process has been executed, or 2) on an intermediate set of of axies after a module has been executed in a process. The modularity in MLDB creation provided by Jeminie therefore allows for a wide range of quality evaluation strategies. The next sections describe the evaluation criteria that we consider for MLDBs created using Jeminie.

3 Taxonomy of evaluation criteria

Here, we propose metrics for the qualitative evaluation of multilingual lexical databases, and give an interpretation for these measures. We propose a classification of MLDB evaluation criteria into four classes, according to their nature.

3.1 Golden-standard-based criteria

In the domain of machine translation systems, an increasingly accepted way to measure the quality of a system is to compare the outputs it produces with a set of reference translations, considered as an approximation of a golden standard (Papineni et al., 2002; hovy et al., 2002). By analogy, one can define a golden standard multilingual lexical database to compare to a database generated by a system such as Jeminie, that both contain axies that link to lexies in the same monolingual databases. Considering that two axies are the same if they contain links to exactly the same lexies, the quality of a machine generated multilingual lexical database

would then be measured with two metrics adapted from machine translation system evaluation (Ahrenberg et al., 2000): recall and precision.

Recall (coverage) is the number of axies that are defined in both the generated database and in the golden standard database, divided by the number of axies in the golden standard.

Precision is the number of axies that are defined in both the generated database and in the golden standard database, divided by the number of axies in the generated database.

However, (Aimelet et al., 1999) highlighted the limits of the golden standard approach, as it is often difficult to manually produce precise reference resources. In the context of the Papillon project, a golden standard multilingual lexical database would deal with nine languages (English, French, German, Japanese, Lao, Thai, Malay, Vietnamese and Chinese), which makes it extremely difficult to produce. Furthermore, since the produced multilingual lexical database in Papillon will define at least 40000 axies, using heterogeneous resources, a comparison with a typical golden standard of only 100 axies seems not relevant. Instead of producing a golden standard for a whole multilingual lexical database, we propose to consider *partial golden standard* that concerns only a part of a MLDB. For instance, a partial golden standard can be produced using a bilingual dictionary that concerns only two languages in the database. Several partial golden standard MLDBs could be produced using several bilingual dictionaries, in order to cover all languages in the multilingual lexical database.

3.2 Structural criteria

Structural evaluation criteria consider the state of links between lexies and axies. We define several general structural criteria:

- CLA_{ave}, the average number of axies linked to each lexie. Here, we consider only lexies that are linked to axies. CLA_{ave} should be 1. If it is > 1, several axies have the same sense, i.e. the produced MLDB is ambiguous. If it is < 1, the produced MLDB may not be precise enough, as it does not cover all the lexies. Actually, we should also consider the standard deviation of that number, because a MLDB would be quite bad if $CLA_{ave} = 2$ for half the lexies and $CLA_{ave} = 0$ for the rest, although the global value of CLA_{ave} is 1.
- for each language, ADL_{lang}, the ratio of the number of axies to the number of lexies in that language. If it is too low, the axies may represent fuzzy acceptions. If it is too high, axies may overlap, i.e. several axies may represent the same acception. Typically, it should be about 1.2 (cf. large MLDB such as EDR - the Electronic Dictionary Research project in Japan). This metrics should be calculated for each language independently, because the number of lexies may significantly vary between two languages, making this metrics irrelevant if calculated using the total number of lexies and axies in a database.
- CAL_{ave}, the average number of lexies of each language linked to each axie. It should be about 1.2. If it is > 1 for a language, axies may represent a fuzzy acception or there is synonymy, as illustrated in figure 3. If it is < 1 for a language, axies may not cover that language precisely. Note that CAL_{ave} may help us locate places in the "axie" set where an axie is refined by one or more axies. Each CAL_{ave} may then be far from CAL_{ave} global, but their average should still be near CAL_{ave} global for the considered set.

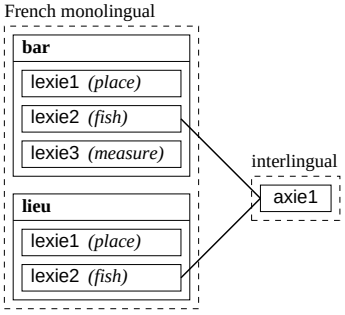


Figure 3: Example of two lexies that are synonym in the same language and linked to the same axie

Such metrics are complementary and can easily be measured, and are among the rare metrics that concern a whole MLDB. They, however, do not help evaluating the quality of links between axies and lexies in terms of semantics.

3.3 Human-based criteria

This class of evaluation criteria is based on the measurement of the number and nature of the *corrections made by a linguist* on a part of a

produced MLDB. For instance, one can measure the ratio of the number of corrections made by a linguist, to the total number of links between the considered axies and lexies. The closer the ratio is to zero, the higher is the quality of the multilingual lexical database. A high correction ratio implies a low MLDB quality.

However, this class of criteria assumes that the produced MLDB are homogeneous. In the context of Papillon, the database will be produced using several techniques and heterogeneous lexical resources, which limits the relevance of such criteria.

This approach is similar to the golden-standard approach described above, although the golden-standard approach is automatic.

3.4 Non-resource-based semantic criteria

In this class, criteria evaluate the quality of the semantics of the links between axies and lexies, and do not rely on additional lexical resources. One of the metrics that we consider is the distance between conceptual vectors of lexies linked to the same axie. A conceptual vector for a lexie is calculated by projecting the concepts associated with this lexie into a vector space, where each dimension corresponds to a leaf concept of a thesaurus (Lafourcade, 2002). The concepts associated with a lexie are identified by analyzing the lexie definition. The lower the distance between the conceptual vectors of two lexies is, the closer are those lexies (word-senses). As a metrics, we therefore consider the average conceptual distance between each pair of lexies linked to the same axie. The lower that value is, the better the MLDB is, in terms of the semantics of the links between axies and lexies. However, a reliable computation of conceptual vectors relies on the availability precise and rich definitions in lexies, and on large lexical resources to compute initial vectors, which are difficult to gather for all languages in practice.

3.5 Discussion

As a more general conceptual framework, we define a classification of evaluation criteria along four *dimensions*, or characteristics:

- *automation:* a criterion is either automatically evaluated, or relies on linguists.
- *scope:* a criterion evaluates either a part of a MLDB, or a whole MLDB.
- *semantics:* a criterion considers either the structure of a MLDB, or the semantics of the links between axies and lexies.
- *resource:* a criterion relies on additional lexical resources, or not.

Multilingual lexical databases such as Papillon can be used in different contexts, e.g. in machine translation systems or in multilingual information retrieval systems. The criteria used for evaluating a multilingual lexical database should be *adapted to the context* in which the database is used. For instance, if a multilingual lexical database is very precise and good at French and Japanese acceptions, but not good at other languages, it should be judged as a good lexical database by users who evaluate a usage of French and Japanese only, but it should be judged as a bad multilingual lexical database globally.

Since the Papillon database generated by Jeminie will not be tied to specific usages, the database production system must not impose predefined evaluation criteria. We propose instead to allow for the use of any criterion at any point in the four dimensions above and for *arbitrary composition of evaluation criteria* to adapt to different contexts. However, since we aim at performing an automatic evaluation, we do not consider human-based criteria, although human evaluation is certainly valid. Our approach is similar to the approach chosen in Jeminie for the creation of axies. We tackle this problem of criteria composition from a software engineering point of view, by using object oriented programming techniques to design and implement modular and reusable *criterion software modules*.

4 Adaptable evaluation system

By analogy with the Jeminie modules that implement algorithms to create axies, we propose a system that allows for the implementation in Java of reusable software modules that implement algorithms to measure MLDB. In this system, we consider that each criterion is implemented as a module. Criterion modules are of a different kind, and are developed differently from Jeminie axie creation modules. As a convention, we define that each criterion module returns a numeric value as the result of a measurement, noted Q_i. The higher that value, the better the evaluated database.

4.1 Axie-creation-related criteria

As the strategy we have chosen in Jeminie is to *combine complementary axie creation modules* to produce axies in a multilingual lexical database, we consider that each axie creation module encapsulates its own quality criterion that it tends to optimize, explicitly or implicitly. Since each module implements an algorithm to *decide whether to create an axie*, we consider that such an algorithm can also be used as a criterion to *decide whether an existing axie is correct*. An axie creation module can not be reused as is as a criterion module, however its decision algorithm can be easily reimplemented in a criterion module. For each algorithm, we define the following four metrics, adapted from (Bédécarrax, 1989):

A_1 the number of *internal adjustments*, i.e. the number of axies that would be created according to the algorithm, and that have actually been created.

A_2 the number of *external adjustments*, i.e. the number of axies that would *not* be created according to the algorithm, and that have actually *not* been created.

E_1 the number of *internal errors*, i.e. the number of axies that would *not* be created according to the algorithm, and that have actually been created.

E_2 the number of *external errors*, i.e. the number of axies that would be created according to the algorithm, and that have actually *not* been created.

For each algorithm, the quality criteria are to maximize $A_1 + A_2$, to minimize $E_1 + E_2$, or to maximize $(A_1 + A_2) - (E_1 + E_2)$.

Resource-based algorithms

For instance, following are the definitions of A_1, A_2, E_1 and E_2 for the axie creation algorithm that uses a bilingual dictionary between languages X and Y:

A_1 the number of pairs of lexies of languages X and Y that are linked to the same axie and which words are mutual translations according to the bilingual dictionary.

A_2 the number of pairs of lexies of languages X and Y that are not linked to the same axie and which words are not mutual translations according to the bilingual dictionary.

E_1 the number of pairs of lexies of languages X and Y that are linked to the same axie and which words are not mutual translations according to the bilingual dictionary.

E_2 the number of pairs of lexies of languages X and Y that are not linked to the same axie and which words are mutual translations according to the bilingual dictionary.

However, resources used by resource-based creation algorithms have a number of entries that is often significantly lower than the number of lexies and axies in a multilingual lexical database. For instance, the number of translation entries in a bilingual dictionary is typically lower than the number of available monolingual acceptions in the source language, because that set of lexies may be constructed by combining a set of rich monolingual dictionaries. For instance, our monolingual database for French contains about 21000 headwords and 45000 lexies extracted from many definition dictionaries such as Hachette, Larousse, etc. Our monolingual database for English contains about 50000 headwords and 90000 lexies extracted from English WordNet 1.7.1. However, the bilingual French-English dictionary that we use is based on the FeM[2] multilingual dictionary, and defines only 15000 French headwords.

lexical database	number of headwords
French monolingual	21000
English monolingual	50000
FeM	15000

Table 1: Comparing the number of entries in monolingual lexical databases with the number of entries in the multilingual lexical database

According to the example above, measuring the number of external adjustments A_2 and internal errors E_1 is therefore not relevant. For example, a criterion can not decide if the words of a French lexie and of an English lexie that are linked together, are translations of each other, since the bilingual dictionary used is not precise enough. We therefore propose a *simplified quality criterion* for resource-based algorithms, that is to maximize $A_1$and to minimize E_2.

[2]French-English-Malay dictionary http://www-clips.imag.fr/geta/services/fem

Vectorial algorithms

This measure can also be adapted to the comparison of the conceptual distance between lexies:

A_1 the number of pairs of lexies that are linked to the same axie and which conceptual vector distance is below a given threshold.

A_2 the number of pairs of lexies that are not linked to the same axie and which conceptual vector distance is above the threshold.

E_1 the number of pairs of lexies that are linked to the same axie and which conceptual vector distance is above the threshold.

E_2 the number of pairs of lexies that are not linked to the same axie and which conceptual vector distance is below the threshold.

This algorithm is not limited by the size of an additional lexical resource, and can decide whether any pair of lexies should be linked or not. It is therefore possible to evaluate A_2 and E_1 in addition to A_1 and E_2.

Synthesis

We specify that the value returned by such axie-creation-related criteria is calculated as $Q_i = A_1 - E_2$ for resource-based criteria, and as $Q_i = (A_1+A_2)-(E_1+E_2)$ for any other axie-creation-related criteria, as those formulas reflect both the number of adjustments and the number of errors.

4.2 Structural criteria

As described above, structural criteria consider the structure of each axie in a whole multilingual lexical database. We propose to implement such algorithms also as modules in our system. For example, we define one criterion module to calculate the following value:

$$Q_i = \frac{1}{0.01 + \left| 1 - \sum_{k=1}^{nblexies} \frac{nblinkedaxies_k}{nblexies} \right|}$$

where $nblexies$ is the total number of lexies in the database, and $nblinkedaxies_k$ is the number of axies linked to a lexie k. Q_i is comprised between 0 and 100.

4.3 Global criteria

A global quality value Q can be calculated as the sum of each quality value measured by each measurement module. The choice of the measurement modules corresponds to a given usage context of the evaluated database, and the positive weight of each metric module in this context is specified as a factor in the sum:

$$Q = \sum_{i=1}^{nbmodules} weight_i \cdot Q_i$$

The objective is to maximize Q. The weight for each module can be chosen to emphasize the importance of selected criteria in the context of evaluation. For instance, when specifically evaluating the quality of axies between French and English lexies, the weight for a bilingual EN-FR dictionary-based criterion module could be higher than the weights for the other criterion modules. In addition, the values returned by different criterion modules are not normalized. It is therefore necessary to adapt the weights to compensate the difference of scale between Q_i values.

5 Evaluation method

One can evaluate the quality of a MLDB after it has been created or enhanced through the execution of an axie creation process by Jeminie. Such a quality measure can be used by linguists to *decide* whether to execute another axie creation process to enhance the quality of the database, or to stop if the database has reached the desired quality. The creation of an axie database is therefore *iterative*, alternating executions of axie creation processes, quality evaluations, and decisions.

It should be noted that the execution of an axie creation process may not always imply a monotonous increase of the measured quality. Since axie creation algorithms may not be mutually coherent, the order of executions of modules, in a process or in several consecutively executed processes, has an impact on the measured global quality. More precisely, the additional resources used by axie creation modules, and/or by quality criteria modules, may contain errors and be mutually incoherent. The execution of a resource-based axie creation module using a resource R_1, can cause a drop of the A_1 value and an increase of the E_2 value measured by a resource-based criterion module using a resource R_2 incoherent with R_1. This may significantly decrease the evaluated global quality. The database may however be actually of a better quality if R_2 has a poor quality and R_1 has a

good quality. This highlights the need for good quality resources for both creating the database and evaluating its quality.

Another problem is that the additional lexical resources used, such as bilingual dictionaries, generally provide information at the level of words, not at the level of senses. It is thus necessary to complement these resource-based axie creation modules, for instance by using vectorial modules. Moreover, it is necessary to develop new algorithms to increase the internal consistence of an axie database, for example one that merges all the axies that link to the same lexie.

6 Example processes

Figure 4 illustrates the two sets of axies created by a process A and a process B to link to lexies retrieved from a French and an English monolingual dictionaries. Process A consists of the execution of only module Mbidict, that uses a bilingual dictionary FR-EN extracted from FeM dictionary and partially illustrated in figure 5. The set of axies produced by process A consists of axie1 to axie7. Process B consists of the execution of the same module Mbidict as in process A, then of a module Mvect that implements a conceptual vector comparison algorithm for filtering some bad links. Process B produces only axie1, axie4, axie5 and axie7. Note that processes A and B were hand-simulated in this example.

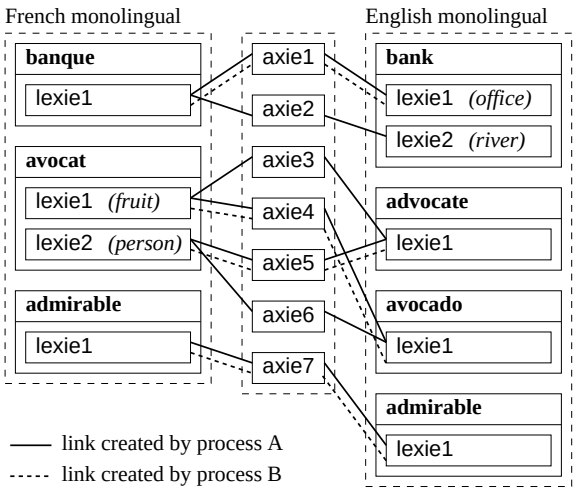


Figure 4: Axies created by processes A and B

The two same criterion modules are used to evaluate both processes: 1) an axie-creation-related criterion module using the same bilingual dictionary as the one used in the axie creation modules in processes, and calculating a Q_{bidict} value, and 2) the structural criterion module described in section 4.2, and calculating a Q_{struct} value. The global evaluated quality value for the set of axies created by each process is:

$$Q = \alpha \cdot Q_{bidict} + \beta \cdot Q_{struct}$$

The actually evaluated values of Q_{bidict} and Q_{struct}, and of Q for several combinations of α and β, are shown in table 2.

Bilingual Dictionary FR–EN (FeM)

Banque (n.f.)	— *Bank (n.)*
Admirable (a.)	— *Admirable (a.)*
Avocat (n.m.)	— *Advocate (n.)*
Avocat (n.m.)	— *Avocado (n.)*

Bilingual Dictionary EN–FR (Le Robert & Collins)

Bank (n.)	— Banque (n.f.)
Bank (n.)	— Rive (n.f.)
Admirable (a.)	— Admirable (a.)
Advocate (n.)	— Avocat (n.m.)
Avocado (n.)	— Avocat (n.m.)

Figure 5: Bilingual dictionaries

	process A	process B
Q_{bidict}	7	1
Q_{struct}	1.76	8.25
Q (α=1, β=1)	8.76	9.25
Q (α=1, β=2)	10.52	17.5
Q (α=2, β=1)	15.76	10.25

Table 2: The results of qualitative evaluations

Axie creation module Mbidict considers only words, but not senses of words. It therefore creates several axies linked to each lexie, some of which are not correct because they do not distinguish between the lexies of a given translation word. In process B, module Mvect is executed to suppress links and axies that are semantically incorrect. The structural quality, as given in Q_{struct}, is therefore better with process B than with process A, and intuitively the global quality has actually increased. However, executing module Mvect reduces the quality from the point of view of a bilingual translation that considers only words and not acceptions, as given in Q_{bidict}.

This illustrates that not all quality criteria should be maximized to attain the best possible quality. Weight factors for each criterion module should be carefully chosen, according to the scale of the values returned by each module, and to the linguistic objectives. For instance, as illustrated in table 2, setting a weight too high for the bilingual translation criterion lets the evaluated global quality decrease, while it has actually increased.

7 Conclusion

This article presents the problem of the automatic creation and evaluation of interlingual multilingual lexical databases (MLDB), in the context of the Papillon project. It describes the Jeminie software system, that we are developing, for the automatic creation of interlingual acceptions (axies). It can adapt to different contexts, e.g. to different lexical resources and different languages, by providing a means to arbitrarily compose axie creation modules.

We have proposed a taxonomy of criteria for the automatic evaluation of a MLDB. One criteria alone is not sufficient to significantly evaluate the quality of a whole database. We therefore propose a method for the arbitrary composition of evaluation criteria, following the same principles as the Jeminie system.

The proposed method will be implemented in a software framework, along with a library of modules that implement a variety of evaluation criteria, and that can be freely composed. This framework will be integrated with Jeminie, in order to allow for the automatic evaluation of a MLDB during its creation.

References

Lars Ahrenberg, Magnus Merkel, Anna Sagvall Hein, and Jorg Tiedemann. 2000. Evaluation of word alignment systems. In *Proceeding of LREC'2000*, pages 1255–1261, Athens, Greece.

Elisabeth Aimelet, Veronika Lux, Corinne Jean, and Frédérique Segond. 1999. WSD evaluation and the looking-glass. In *Proceedings of TALN'1999*, Cargèse, France.

Chantal Bédécarrax. 1989. *Classification automatique en analyse relationnelle : la quadridécomposition et ses applications.* thesis, Université Paris 6.

Etienne Blanc. 1999. PARAX-UNL: A large scale hypertextual multilingual lexical database. In *Proceedings of 5th Natural Language Processing Pacific Rim Symposium*, pages 507–510, Beijing. Tsinghua University Press.

Eduard hovy, Margaret King, and Andrei Popescu-Belis. 2002. Principles of context-based machine translation evaluation. *Machine Translation*, 17(1):43–75.

Mathieu Lafourcade. 2002. Automatically populating acception lexical databases through bilingual dictionaries and conceptual vectors. In *Papillon'2002 Seminar*, Tokyo, Japan.

Igor Mel'cuk, André Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire.* Duculot, Louvain-la-Neuve.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceeding of ACL'2002*, pages 311–318, Philadelphia.

Alain Polguère. 2000. Towards a theoretically motivated general public dictionary of semantic derivations and collocations for French. In *Proceedings of EURALEX'2000*, pages 517–527, Stuttgart.

Gilles Sérasset. 1994. Interlingual lexical organisation for multilingual lexical database in NADIA. In *Proceedings of COLING'94*, volume 1/2, pages 278–282, Kyoto, Japan.

Aree Teeraparbseree. 2003. Jeminie: A flexible system for the automatic creation of interlingual database. In *Papillon'2003 Seminar*, Sapporo, Japan.

Automatic Construction of a Transfer Dictionary Considering Directionality

Kyonghee Paik, Satoshi Shirai* and **Hiromi Nakaiwa**
{kyonghee.paik,hiromi.nakaiwa}@atr.jp * sat@fw.ipsj.or.jp

ATR Spoken Language Translation Laboratories
2-2-2, Keihanna Science City Kyoto, Japan 619-0288

*NTT Advanced Technology Corporation
12-1, Ekimaehoncho, Kawasaki-ku, Kawasaki-shi, Japan 210-0007

Abstract

In this paper, we show how to construct a transfer dictionary automatically. Dictionary construction, one of the most difficult tasks in developing a machine translation system, is expensive. To avoid this problem, we investigate how we build a dictionary using existing linguistic resources. Our algorithm can be applied to any language pairs, but for the present we focus on building a Korean-to-Japanese dictionary using English as a pivot. We attempt three ways of automatic construction to corroborate the effect of the directionality of dictionaries. First, we introduce "one-time look up"method using a Korean-to-English and a Japanese-to-English dictionary. Second, we show a method using "overlapping constraint" with a Korean-to-English dictionary and an English-to-Japanese dictionary. Third, we consider another alternative method rarely used for building a dictionary: an English-to-Korean dictionary and English-to-Japanese dictionary. We found that the first method is the most effective and the best result can be obtained from combining the three methods.

1 Introduction

There are many ways of dictionary building. For machine translation, a bilingual transfer dictionary is a most important resource. An interesting approach is the *Papillon Project* that focuses on building a multilingual lexical data base to construct large, detailed and principled dictionaries (Boitet et al., 2002). The main source of multilingual dictionaries is monolingual dictionaries. Each monolingual dictionary is connected to interlingual links. To make this possible, we need many contributors, experts and the donated data. One of the studies related to the *Papillon Project* tried to link the words using definitions between English and French, but the method can be extended to other language pairs (Lafourcade, 2002). Other research that focuses on the automatic building of bilingual dictionaries include Tanaka and Umemura (1994), Shirai and Yamamoto (2001), Shirai et al. (2001), Bond et al. (2001), and Paik et al. (2001).

Our main concern is automatically building a bilingual dictionary, especially with different combinations of dictionaries. None of the research on building dictionaries seriously considers the characteristics of dictionaries. A dictionary has a peculiar characteristic according to its directionality. For example, we use a Japanese-to-English (henceforth, J⇒E) dictionary mainly used by Japanese often when they write or speak in English. Naturally, in this situation, a Japanese person knows the meaning of the Japanese word that s/he wants to translate into English. Therefore, an explanation for the word is not necessary, except for the words whose concept is hard to translate with a single word. Part-of-speech (henceforth POS) information is also secondary for a Japanese person when looking up the meaning of the corresponding equivalent to the Japanese word.

On the other hand, an English-to-Japanese (henceforth E⇒J) dictionary is basically used from a Japanese point of view to discover the meaning of an English word, how it is used and so on. Therefore, explanatory descriptions, example sentences, and such grammatical information as POS are all important. As shown in (2), a long explanation is used to describe the meaning of *tango*, its POS and such grammatical information as singular or plural. Also, an E⇒J dictionary includes the word in plenty of

* Some of this research was done while at ATR.

examples, comparing to a J⇒E dictionary. The following examples clearly show the difference.

(1) J⇒E: タンゴ:《dance》the tango 《〜**s**》

(2) E⇒J: tan · go /(**n**. **pl** 〜**s**)
タンゴ:**a**. もと中央アフリカの原住民の舞踏..etc. (trans. tango "a dance of Central African aboriginals,...etc.")**b**. その曲(trans. "its music")**Vi** タンゴを踊る("to dance the tango").

In this paper, we evaluate the effects that occur when we use different combinations of dictionaries and merge them in different ways.

2 Conventional Methods and Problems

The basic method of generating a bilingual dictionary through an intermediate language was proposed by Tanaka and Umemura (1994). They automatically constructed a Japanese-French dictionary with English as an intermediate language and manually checked the extracted results. In this sense, their method is not completely automatic. They looked up English translations for Japanese words, and then French translations of these English translations. Then, for each French word, they looked up all of its English translations. After that, they counted the number of shared English translations (**one-time inverse consultation**). This was extended to "two-time inverse consultation". They looked up all the Japanese translations of all the English translations of a given French word and counted how many times the Japanese word appears. They reported that "comparing the generated dictionary with published dictionaries showed that data obtained are useful for revising and supplementing the vocabulary of existing dictionaries." Their method shows the basic method of building a dictionary using English as an intermediate language. We applied and extended their method in automatic dictionary building especially considering the directionality of dictionaries.

Tanaka and Umemura (1994) used four dictionaries in two directions (J⇒E, E⇒J, F⇒E and E⇒F). They first harmonized the dictionaries by combining the J⇒E and E⇒J into a single J⇔E and the F⇒E and E⇒F into a harmonized F⇔E dictionary. We followed their basic method without harmonizing the dictionaries to emphasize the influence of directionality.

In general, foreign word entries in a bilingual dictionary attempt to cover the entire vocabulary of the foreign language. However, foreign words that do not correspond to one's mother tongue are not recorded in a bilingual dictionary from one's mother tongue to the foreign language (Hartmann, 1983). A long explanatory phrase is replaced with a word that often does not perfectly correspond to the original.

On the other hand, most of the index words from a foreign language to a mother tongue include many expository definitions or explanations that focus on usage. Such syntactic information as POS and number as well as example sentences are rich compared with a dictionary from mother tongue to a foreign language. These characteristics should be considered when building a dictionary automatically.

Bond et al. (2001) showed how semantic classes can be used along with an intermediate language to create a Japanese-to-Malay dictionary. They used semantic classes to rank translation equivalents so that word pairs with compatible semantic classes are chosen automatically as well as using English to link pairs. However, we cannot use this method for languages with poor language resources, in this case semantic ontology. Paik et al. (2001) improved the method to generate a Korean-to-Japanese (henceforth K⇒J) dictionary using multi-pivot criterion. They showed that it is useful to build dictionaries using appropriate multi-pivots. In this case, English is the intermediate language and shared Chinese characters between Korean and Japanese are used as pivots.

However, none of the above methods considered the directionality of the dictionaries in their experiments. We ran three experiments to emphasize the effects of directionality.[1] There are many approaches to building a dictionary. But our focus will be on the generality of building any pair of dictionaries automatically using English as a pivot. In addition, we want to confirm various directionalities between a mother tongue and a foreign language.

[1]The first two experiments were reported in Shirai and Yamamoto (2001) and Shirai et al. (2001). We present new evaluations in this paper.

3 Proposed Method

We introduce three ways of constructing a K⇒J dictionary. First, we construct a K⇒J dictionary using a K⇒E dictionary and a J⇒E. Second, we show another way of constructing a K⇒J dictionary using an K⇒E dictionary and an E⇒J dictionary. Third, we use a novel way of dictionary building using an E⇒K and E⇒J to build a K⇒J dictionary. However, our method is not limited to building a K⇒J dictionary but can be extended to any other language pairs so long as X-to-English or English-to-X dictionaries exist. These three methods will cope with making dictionaries using any combination.

We assume that the following conditions hold when building a bilingual dictionary: (1) Both the source language and the target language cannot be understood (to build a dictionary of unknown language pairs); (2) Various lexical information of the intermediate language (English) is accessible. (3) Limited information about the source and target language may be accessible.

3.1 Lexical Resources

Our method can be extended to any other language pairs if there are X-to-English and English-to-X dictionaries. It means that there are four possible combinations such as i) X-to-English and Y-to-English, ii) X-to-English and English-to-Y, iii) English-to-X and Y-to-English and iv) English-to-X and English-to-Y to build a X-to-Y dictionary. We tested i), ii) and iv) in this paper and we used the following dictionaries in our experiment:

Type	# Entries	Dictionary
J⇒E	28,310	New Anchor [2]
E⇒J	52,369	Super Anchor[3]
K⇒E	50,826	Yahoo K⇒E [4]
E⇒K	84,758	Yahoo E⇒K [4]

3.2 Linking K⇒E and J⇒E

Our method is based upon a **one-time inverse consultation** of Tanaka and Umemura (1994)(See Section 2.) to judge the word correspondences of Korean and Japanese.

Lexical Resources used here is a K⇒E dictionary (50,826 entries) and a J⇒E dictionary (28,310 entries). There is a big difference in the number of entries between the two dictionaries. This will affect the total number of extracted words.

For Evaluation, we use a similarity score S_1 for a Japanese word j and a Korean word k is given in Equation (1), where $E(w)$ is the set of English translations of w. This is equivalent to the Dice coefficient. The extracted word pairs and the score are evaluated by a human to keep the accuracy at approximately 90%.

$$S_1(j,k) = \frac{2 \times |E(j) \cap E(k)|}{|E(j)| + |E(k)|} \qquad (1)$$

The most successful case is when all the English words in the middle are shared by K⇒E and J⇒E. Figure 1 shows how the link is realized and the similarity scores are shown in Table 1. The similarity score shows how many English words are shared by the two dictionaries: the higher the score, the higher possibility of successful linking. However, as Table 1 shows, we have to sort out the inappropriately matched pairs by comparing the S_1 score of equation (1) against a threshold τ. The threshold allows us to exclude unfavorable results. For example, for words having one shared English translation equivalent, we have to discard the group (3) in Table 1.

When the words translated from English match completely, the accuracy is high. And if the number of shared English translated words ($|E(J) \cap E(K)|$) is high, then we get a high possibility of accurate matching of Korean and Japanese. However, accuracy deteriorates when the number of the shared English translated words (shown by the threshold) decreases as in (2) and (3) of Table 1. We solved this problem by varying the threshold according to the number of shared English equivalents. The value of the threshold τ was determined experimentally to achieve an accuracy rate of 90%.

Result: Linking through English gives a total of 175,618 Korean-Japanese combinations. To make these combinations, 28,479 entries out of 50,826 from the K⇒E dictionary and 17,687 entries out of 28,310 from the J⇒E dictionary are used. As a result, we can extract 25,703 estimated good matches with an accuracy of 90%.

[2](Yamagishi et al., 1997) [3] (Yamagishi and Gunji, 1991) [4] `http://kr.engdic.yahoo.com`

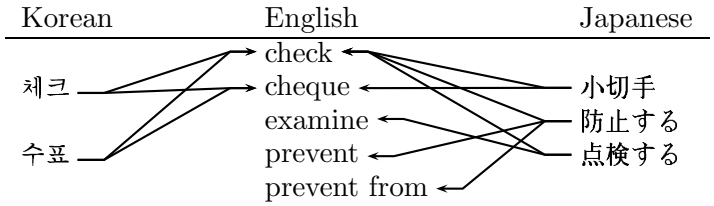


Figure 1: Linking through English translation equivalents (K⇒E, J⇒E)

	Shared Eng.	τ	Korean ⇒ English	Japanese ⇒ English
(1)	2	1.000	(체크check;cheque)	(小切手check;cheque)
	2	1.000	(수표 check;cheque)	(小切手check;cheque)
(2)	1	.667	(체크 check;cheque)	(照合check)
(3)	1	.500	(체크check;cheque)	(点検するcheck;examine)
	1	.400	(체크check;cheque)	(防止するprevent from;prevent;check)
	1	.333	(수표check;cheque)	(預けるleave;deposit;check;entrust)

Table 1: Example of linking through English translations

Shared Eng[5]	Extracted	τ	Good matches
7	1	0	1
6	1	0	1
5	16	0	16
4	165	0	165
3	1,325	0.4	1,206
2	12,037	0.5	7,401
1	161,863	0.667	16,790
Total	175,408		25,580

Table 2: Matching words by K⇒E +J⇒E

3.3 Linking K⇒E and E⇒J

Method: We investigated how to improve the extraction rate of equivalent pairs using an **overlapping constraint** method here. To extract Korean-Japanese word pairs, we searched consecutively through a K⇒E dictionary and then an E⇒J dictionary. We take English sets corresponding to Korean words from a Korean-English dictionary and Japanese translation sets for each English words from an E⇒J dictionary. The overlap similarity score S_2 for a Japanese word j and a Korean word k is given in Equation (2), where $E(w)$ is the set of English translations of w and $J(E)$ is the bag of Japanese translations of all translations of E.

$$S_2(j,k) = |j|, j \in J(E(k)), \qquad (2)$$

After that, we test the narrowing down of translation pairs by extracting the overlapped words in the Japanese translation sets. See Figure 2.

Lexical Resources: We used a K⇒E dictionary (50,826 entries), the same as the one used in section 3.2 and a E⇒J dictionary (52,369 entries). Compared to the resources used in our first method, the number of entries are well balanced.

Evaluation: After extracting the overlapped words in the Japanese translation sets, the words were evaluated by humans. The main evaluation was to check the correlation between the overlaps and the matches of Korean and Japanese word pairs. Table 3 shows the overlapped number of shared English words and the number of index words of the K⇒E dictionary.

Overlaps	Num of entries in K⇒E
4 or more	1,286
3	3,097
2	13,309
1-to-1 match	1,315
Subtotal	19,007
Other match	8,832
No Match	22,987
Total	50,826

Table 3: The number of entries in K⇒E dictionary according to overlapped English words

Result: Entries with a 1-to-1 match have $|E(K)| = |E(J)| = 1$. These are generally good matches (90%). If more than two overlaps occur, then the accuracy matching rate is as high as 84.0%. It means that the number of useful entries is the sum of the 1-to-1 matches and 2 or more overlaps: 19,007 (37.4% of the K⇒E entries) with 87% accuracy. However, using K⇒E and E⇒J there is a problem of polysemy in English words. For example, *clean* has two different POSs, adjective and verb in a K⇒E dictio-

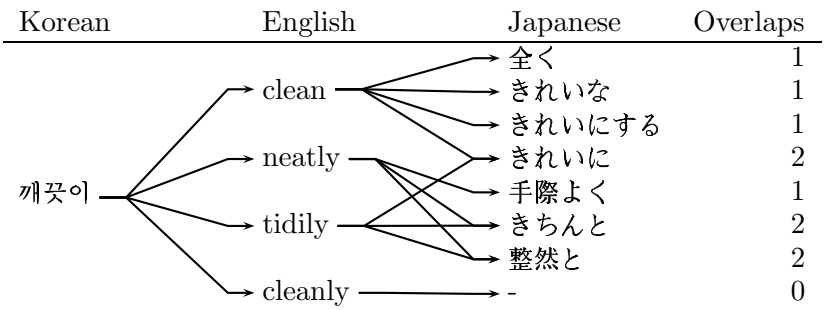


Figure 2: Overlapping Translation equivalents (K⇒E, E⇒J)

nary. Unfortunately, this information cannot be used effectively due to the lack of POS in K⇒E when linking them to a E⇒K dictionary. On the other hand, *clean* using E⇒J can be translated into either きれいな, an adjective or きれいにする, a verb. This makes the range of overlap score widely distributed as shown in Figure 2. This is the reason using K⇒E and E⇒J is not as good as using K⇒E and J⇒E. We will discuss this more in section 4.

3.4 Linking E⇒K and E⇒J

As we have discussed in earlier sections, the characteristics of dictionaries differ according to their directionality. In this section, we introduce a novel method of matching translation equivalents of Korean and Japanese. From the Korean speaker's point of view, the E⇒K dictionary covers all English words, includes explanatory equivalents, and example sentences showing usage. The same thing is true for the E⇒J dictionary from a Japanese speaker's point of view. In this respect, we expect that the result of extraction is not as effective as the other combinations such as K⇒E +J⇒E and K⇒E +E⇒J. On the other hand, we think that there must be other ways to exploit explanatory equivalents and example sentences.

Method: First, we linked all the Korean and Japanese words if there is any shared English words. Then, we sorted them according to POSs to avoid the polysemous problem of POS. The left hand side of Figure 3 shows how we link Korean and Japanese pairs.

Lexical Resources: We used a E⇒K dictionary (84,758 entries) and a E⇒J dictionary (52,369 entries). Both of the dictionaries have many more entries than the ones used in the previous two methods.

Evaluation: We use similarity score S_3 in Equation (3) as a threshold which is used to extract good matches.

$$S_3(k,j) = \frac{|K(E(k) \cap E(j))| + |J(E(k) \cap E(j))|}{|E(k) \cap E(j)|} \tag{3}$$

$K(W)$: bag of Korean translations of set W
$J(W)$: bag of Japanese translations of set W
$E(w)$: set of English translations of word w

$|K(E)|$ means the number of Korean translation equivalents, and$|J(E)|$ means the number of Japanese translation equivalents. The sum of the numbers is divided by the number of intermediate English words. It is used to reduce the polysemous problem of English words. It is because it is hard to decide which translation is appropriate, if an English word has too many translation equivalents in Korean and Japanese. The value of threshold (S_3) is shown in Table 4. We vary the threshold according to $N = |E(j) \cap E(k)|$ to maximize the number of successful matches experimentally. N represents the number of intermediate English words. For N=1, we only count one-to-one matches, which means one Korean and one Japanese are matched through only one English. The following are examples of being counted when N is 1-to-1: e.g. 자기 암시-autosuggestion(n.)- 自己暗示, 당구(용)의-billiard(a.)-玉突きの, etc. We may lose many matching pairs by this threshold, but the accuracy rate for 1-to-1 is very high (96.5%). To save other matches when N=1, we need to examine further. In our experiment, 귀여운⇔ 愛らしいis rejected because lovely has two Korean translations and two Japanese translations; the match 귀여운⇔ 愛らしいis not 1-to-1. We postpone this part to further research.

N	Extracted	Matched	Good	S_3	Extracted	Matched	Good
24-6	438	422	96.3%	any	438	422	96.3%
5	313	301	96.2%	≤ 35	302	293	97.0%
4	790	698	88.3%	≤ 25	661	601	90.9%
3	2,432	1,960	80.7%	≤ 10	634	586	92.4%
2	12,862	(6,784)	(52.8%)	≤ 10	3,613	(3,150)	(87.2%)
$*$1[-to-1]	4,712	(4,547)	(96.5%)	2	4,712	(4,547)	(96.5%)
	21,547	(14,712)	(68.3%)		10,360	(9,599)	(92.7%)

Table 4: Summary of matching words by E⇒K and E⇒J
N: Number of total English translation equivalents
$*$: We only count word pairs under the condition of 1-to-1 match.

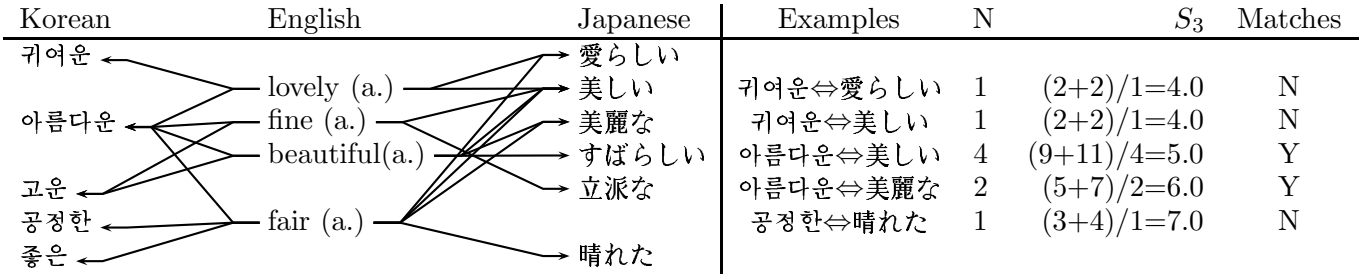


Figure 3: An example of matching E⇒K and E⇒J

Result: Table 4 shows the extracted 21,564 pairs of Korean and Japanese words. On average, 14,712 pairs match with a 68.3% success rate. The numbers in parentheses are estimated.

As expected, by setting this threshold we get fewer extracted words such as 10,360 words as shown in Table 4. However, the accuracy of the matched word pairs averages 92.7%.

Comparison: To compare the three methods, we randomly chose 100 Korean words from a K⇒J dictionary[6] which could be matched through all three methods. The number of extracted matches was 28 using K⇒E and J⇒E, 34 using K⇒E and E⇒J, and 13 using E⇒K and E⇒J. For K⇒E and E⇒J method, 21 out of 34 K⇒J pairs were found only in K⇒E and E⇒J method but not in K⇒E and J⇒E method. Among the 21 new K⇒J word pairs, only one pair is an error (not a good match). One new pair was found in E⇒K and E⇒J method. Therefore, combining all three methods gave 49 (28+20+1) different K⇒J pairs, a better result than any single method. These results are shown in Table 5. Clearly the dictionaries used greatly affect the number of matches. The number of matches could be improved by considering English derived forms (e.g. matching *confirmation* with *confirm*).

	K⇒E +J⇒E	K⇒E +E⇒J	E⇒K +E⇒J
Total	28	34	13
Good	28	33	10
Error	0	1	3

Table 5: Comparison of the Proposed Methods

4 Discussion

We have shown the results of different matching metrics for different dictionary directions. Directionality is an important matter for building dictionaries automatically. In a K⇒E (or J⇒E) dictionary an index word contains non-conjugated forms whereas an index word in E⇒K (or E⇒J) dictionary contains POS and conjugated forms. Therefore we expect the combination of K⇒E and J⇒E to be better than K⇒E and E⇒J since we can avoid the mismatch of POS.

On the other hand, a dictionary E⇒K or E⇒J contains less uniform information such as long expository terms, grammatical explanations and example sentences. Especially, POS is far more detailed than the dictionaries of the

[6]We used **Korean-Japanese dictionary** (Shogakukan: 1993) for the sampling that includes 110,000 entries, many of which are used infrequently.

other direction. These all contribute to fewer good matching words.

As for the better result using K⇒E and J⇒E, we cannot overlook language similarity: Korean and Japanese are very similar with respect to their vocabularies and grammars. This must have result in sharing relatively more appropriate English translations and further matching more appropriate Korean and Japanese translation equivalents.

In the combination of K⇒E and E⇒J, the common English translations are reduced due to the characteristics of K⇒E and E⇒J. A K⇒E dictionary from the Korean speaker's point of view tends to have relatively simple English equivalents and normally POS is not shown. On the other hand, an E⇒J dictionary shows such complicated equivalents as explanation of the entry **a**, a piece of translation equivalent **b** and grammatical information as shown in (2) in Section 1. Therefore, it is natural that the matching rate is far less than the combination of K⇒E and J⇒E. Considering the size of dictionaries used in K⇒E and J⇒E (estimated maximum matches: 28,310 K⇒J pairs) and the one used in K⇒E and E⇒J (estimated maximum matches: 50,826 K⇒J pairs), we extrapolate from Table 5 that the method using K⇒E and J⇒E is better than the method using K⇒E and E⇒J.

We concluded that: K⇒E + J⇒E outperforms K⇒E + E⇒J which outperforms E⇒K + E⇒J. The following briefly summarizes the three methods.

- K⇒E + J⇒E:
  - Equal characteristics of the dictionaries
  - The meaning of the registered words tends to be translated to a typical, core meaning in English
  - Synergy effect: Korean and Japanese are very similar, leading to more matching.
- K⇒E + E⇒J:
  - The combination of different characteristics of dictionaries makes automatic matching less successful.
  - A core meaning is extended to a peripheral meaning at the stage of looking up E⇒J. (See Figure 2.)
- E⇒K + E⇒J:
  - There are far fewer matches.
  - We can take advantage of example sentences, expository terms, and explanations to extract functional words.
  - We can improve accuracy by including English POS data.

Even though we expected that the combination of dictionaries between E⇒K and E⇒J will not provide a good result, it is worthwhile to know limits. After analyzing all of the result, we found that there is the effect of dictionary directionality. Also, we confirm that if we can use all the methods and combine them, we will get the best result since the output of the three dictionary combinations do not completely overlap.

Future Work

Our goal is not restricted to making a Korean-Japanese dictionary, but can be extended to any language pair. We assume that we do not know the source and target languages so well that it is not easy to match just the content words. Instead, we need to match automatically any kind of entries, even such functional words as particles, suffixes and prefixes. We think that it is best to extract these functional words by taking advantage of the characteristics of the E⇒K and E⇒J dictionaries. For example, one of the merits of using E⇒K and E⇒J is that we can get conjugated forms such as the Korean adjective **아름다운** which matches the English adjective **beautiful**; it is normally not registered in a K⇒E dictionary because **아름다운** is an adjective conjugated form of the root **아름답다**. Only the root forms are registered in an X-to-English dictionary. Also for verbs, we can get non finite forms using E⇒K and E⇒J dictionaries. As index word, the non-conjugated forms are registered in a J⇒E dictionary such as **きれいだ** meaning *beautiful* or *clean*. However, by using E⇒J, we can get conjugated forms such as **きれいに**, **きれいな** and so forth. Registering all conjugated forms in a dictionary simplifies the development of a machine translation system and further second language acquisition.

The direction from English-to-X contains a lot of example sentences. So far, the idea of using example sentences and idiomatic phrases for dictionary construction has not been adopted. To check the possibility of extracting functional words, we extracted example sentences and idiomatic phrases from E⇒J and E⇒K dictionaries based upon the number of shared English words and look into the feasibility of using them to extract functional words.

We extracted a total of 1,033 paraphrasing sentence pairs between Korean and Japanese with five or more shared English words. Among them, 465 sentences (45%) matched all the English exactly (=), and 373 sentences (36.1%) almost (≈) matched. We give examples below:

= (10) "as for me, give me liberty or give me death." 私としては自由が得られなければ死んだほうがましだ.
"as for me, give me liberty or give me death." 나에게는 자유가 아니면 죽음을 달라.

≈ (8) "he is taller than any other boy in the class." 彼はクラスのだれよりも背が高い.
"Tom is taller than any other boy in his class." 톰은 반에서 누구보다도 키가 크다.
(extracted from E⇒K and E⇒J)

The numbers in parentheses in the above examples represent how many English words are shared between E⇒K and E⇒J. Using these paraphrasing sentences we will examine the effective way of extracting functional words.

Finally we would like to apply our method to open source dictionaries, in particular EDICT (J⇒E, Breen (1995)) and `engdic` (E⇒K, Paik and Bond (2003)). This would make the results available to everyone, so that they can be used in comparative evaluation or further research.

5 Conclusion

We have shown three major combination of dictionaries to build dictionaries. These methods can be applied to any pairs of language; we used a K⇒E dictionary, a J⇒E, an E⇒K dictionary and an E⇒J to build a K⇒J dictionary using English as a pivot.

We applied three different methods according to different combination of dictionaries. First, a one-time look up method (Tanaka and Umemura, 1994) is tried using K⇒E and J⇒E. Second, an overlapping constraint method in one direction is applied using K⇒E and E⇒J. Finally, a novel combination for building a dictionary is attempted using E⇒K and E⇒J. We found that the best result is obtained by the first method. However, by combining all methods we can extract far more entries since the results from the three method do not overlap. Our result shows that 60% of word pairs in the second method are not found in the first or the third method. For the third method (using E⇒K and E⇒J), we could not extract as many matched pairs, but it is potentially useful for extracting conjugated forms and functional words.

Acknowledgments
This research was supported in part by the Ministry of Public Management, Home Affairs, Posts and Telecommunications. We would also like to thank Francis Bond for his comments and discussion.

References

Christian Boitet, Mathieu Mangeot, and Gilles Sérasset. 2002. The Papillon Project: cooperatively building a multilingual lexical data-base to derive open source dictionaries and lexicons. *The 2nd Workshop NLPXML-2002*, pages 93–96, Taipei, Taiwan.

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and construction of a machine-tractable Japanese-Malay dictionary. In *MT Summit VIII*, pages 53–58, Santiago de Compostela, Spain.

Jim Breen. 1995. Building an electronic Japanese-English dictionary. Japanese Studies Association of Australia Conference.

Reinhard Rudolf-Karl Hartmann. 1983. *Lexicography: Principles and Practice*. Academic Press.

Mathieu Lafourcade. 2002. Automatically populating acception lexical database through bilingual dictionaries and conceptual vectors. In *Papillon 2002 Seminar(CD-Rom)*, Tokyo, Japan.

Kyonghee Paik and Francis Bond. 2003. Enhancing an English and Korean dictionary. In *Papillon-2003*, pages CD–rom paper, Sapporo, Japan.

Kyonghee Paik, Francis Bond, and Satoshi Shirai. 2001. Using multiple pivots to align Korean and Japanese lexical resources. In *NLPRS-2001*, pages 63–70, Tokyo, Japan.

Satoshi Shirai and Kazuhide Yamamoto. 2001. Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *ICCPOL-2001*, pages 174–179, Seoul.

Satoshi Shirai, Kazuhide Yamamoto, and Kyonghee Paik. 2001. Overlapping constraints of two step selection to generate a transfer dictionary. In *ICSP-2001*, pages 731–736, Taejon, Korea.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *COLING-94*, pages 297–303, Kyoto.

Katsuei Yamagishi and Toshio Gunji, editors. 1991. *The New Anchor Japanese-English dictionary*. Gakken.

Katsuei Yamagishi, Tokumi Kodama, and Chiaki Kaise, editors. 1997. *Super Anchor English-Japanese dictionary*. Gakken.

Building and sharing multilingual speech resources, using ERIM generic platforms

Georges FAFIOTTE
GETA, CLIPS, IMAG-campus (UJF, Grenoble 1 Univ.)
385 rue de la Bibliothèque, BP 53
F-38041 Grenoble cedex 9
France
georges.fafiotte@imag.fr

Abstract

In the framework of projects ChinFaDial and ERIM we have developed in recent years several platforms allowing to handle various aspects of bilingual spoken dialogues on the web —mainly, spontaneous speech corpus collection through distant human interpreting. Current development of the core ERIM-Interp and ERIM-Collect platforms now includes multimodal user interaction, integration of some machine aids (such as speech turn logs through speech recognition, or tentatively speech machine translation, both based on server-grounded market products), and next, online aids to speakers and/or interpreters.

First collected data should be made available on the web in fall 2004 (DistribDial) along with, as soon as available, a robust version of the collecting platform, in order to promote collaborative building, and sharing, of "raw" unannotated multilingual speech corpora.

A variant of the ERIM environment is to extend to distant *e*-training in interpreting, possibly creating situations which should in turn, in our view, foster larger-scale data collection and sharing in open access mode.

Keywords

Bilingual speech corpora, collaborative corpus collection, spontaneous dialogues, Web-based interpreting, multilingual communication, open-access resources, resource mutualization.

Introduction

Ongoing burst in the development of both portable telecommunications tools open to Internet transactions, and videoconferencing means, is creating rapid expansion of teleservicing and telebusiness applications with spontaneous dialogue, information inquiry, distant negotiation, etc. Multilingualism, now in spoken transaction as it has been in written one, appears as a key issue in distant communication, with sensitive questions, both in supporting the diversity of the native or origin language of conversing users (particularly within the opening European economic area), and in bringing some kind of balance between main "linguae francae" (common languages). Thus new stakes arise in enhancing distant web-based on-line interpreting services.

Meanwhile, Speech Machine Translation (SMT) steadily takes steps towards style spontaneity and multilingualism. In this context though, we face a notorious lack of large open-access corpora of bilingual spoken dialogues.

This led us to study, to model and propose a set of generic platforms, aiming at enhancing distant multilingual multimodal oral communication with full recording and collecting facilities, also addressing expectations from the MT systems engineering community.

The paper first looks over project motivation, then introduces the interpreting and collecting platforms presently available in the ERIM family, with current variants. It then reports on their first use in collecting domain-oriented spontaneously spoken French-Chinese dialogues. Finally we present ongoing or planned development, advocating for collaborative building and voluntary sharing of resulting multilingual resources.

1. Motivations, early prototyping

1.1 Developing multilingual linguistic resources

It is widely recognized that realistic and large corpora are key resources for building Speech Recognition (SR) and Speech MT systems. If the Web has recently been put to use as the largest possible corpus, modeling casual spontaneous spoken language requires transcribed speech corpora of hundreds of hours.

Speech translation systems thus need large parallel translation corpora of transcribed and aligned spontaneous utterances in dialogue context, ideally with complete sets of parse trees. However, few such corpora have been developed (by NEC, ATR

and a few others), and these are not publicly available. Why not? Because these corpora are very expensive to transcribe once collected, and to annotate. After so much time has been spent in compiling a corpus, giving it away seems unreasonable.
Besides, a future research objective is to use collected corpora for studying and modeling real life spontaneous spoken language and dialogues, and possibly to investigate if and how specific linguistic traits can be expected depending on specific dialogue situations, translation process settings, or various multimodal interaction means.
For instance, two speakers in a bilingual dialogue may hear one another's original speech or not, they may use video or fixed images, etc. Their linguistic behavior is expected to vary accordingly: the number of clarification sub-dialogues may vary; third person use or indirect speech may be used more in the presence of a speech translation system than with a human interpreter; the use of deictic and anaphoric elements may turn out to depend on the use of visible markable objects on whiteboards, maps, images.
With these considerations in mind, we thus endeavoured to propose open-acces corpus resources —and therefore open-access collecting resources—, in order to ease collaborative building of "raw" unannotated multilingual translated speech corpora, likely taking advantage of new web-based interpreting situations or scenarios.

1.2 Enhancing multilingual communication on the Web

Some companies have already developed proprietary network-oriented interpreter's cubicles, which are the counterparts of existing fixed installations for interpreting in multilingual meetings (for example at the UN or EU). However, the associated code is not available for research.
Furthermore, our typical scenario is somewhat different from that of classical interpreting, where interpreters are available for the entire duration of the conversations. We rather allow two situations:

- "conference call": speakers establish a schedule, and book a time slot with an interpreter,
- "on demand interpretation": interlocutors try to converse using whatever knowledge they may have of their interlocutor's language, or of a third common language. When the language barrier impedes communication, they ask an available interpreter to jump in to help.

Apart from these practical motivations, we also wish to conduct experimental studies on the effect of combining multimodal resources on bilingual or multilingual conversations. Thus, full recording facilities were required anyhow.

1.3 Pre-ERIM platforms

Other studies of human "consecutive" interpretation have employed multimodal Wizard of Oz platforms (e.g. the EMMI plateform, that we experienced at ATR-ITL for bilingual pilot-experiments [Fafiotte & Boitet, 1994] [Loken-Kim & al., 1994]), or monolingual multi-Wizard architectures have been modelled in a multimodal setting (NEIMO [Coutaz & al., 1996]). Thus our first objective was to produce a simulator of automatic speech translation systems in the same spirit, to gain experience and collect data.
We first built prototypes of a Speech MT Wizard of Oz simulator, Sim* [Fafiotte & Zhai, 1999] (to be read as "Sim-Star", since being a parallel platform to the C-STAR II CLIPS environment). They were designed to run on the Internet, and were originally used on the intranet of CLIPS-GETA. Network-based communications were handled by a client-server communication module developed in Tcl/Tk. Participants could see and hear each other and share an electronic whiteboard, using MBone resources.
The idea of using Wizard of Oz techniques in this context proved quite impractical, and thus was abandoned. Even if an acoustic filter was used to deform the interpreter's voice, participants perceived that a human was speaking. In the end, we realized that, even for true automatic high quality interpretation, there actually might well be a real human "warm body" in the loop anyway. Thus a realistic design for online interpretation could integrate both human and machine interpretation for "partially automatic" Speech MT.
The successive ERIM platforms have been implemented on this basis, in parallel at CLIPS with integrating the French language into multilingual Speech Machine Translation within C-STAR and NESPOLE! international projects. ERIM stands in French for Network-based Environment for Multimodal Interpreting.

2. Distant human interpreting, as a collecting scheme for multilingual spoken dialogues

2.1 Context

At CLIPS-GETA, one of the ultimate research goals in Speech MT is to build systems for automatic or partially automatic Speech Interpretation (i.e. "synergic" user-aided translation of speech). Much progress has been made in this area over the past twelve years. NEC produced the first speech translation demo in September 1992, within the tourist domain, but the most widely known coordinated research efforts to date include the C-STAR projects (now a 7-language

international Consortium for Speech Translation Advanced Research) [http://www.c-star.org], the European NESPOLE! project [http://nespole.itc.it], the German Verbmobil [http://verbmobil.dfki.de] project, the US DARPA Communicator program with the Galaxy Communicator Software Infrastructure [http://fofoca.mitre.org/doc.html] [http://www.darpa.mil/ito/research/com/index.html] [http://www.sls.lcs.mit.edu/sls/whatwedo/architecture.html].
All have demonstrated platforms enhancing spontaneous speech processing in multilingual person-person or person-system communication, always in restricted domains. CLIPS is firmly involved in this action, while being in charge for integrating the French language in the C-STAR and NESPOLE! environments.
At the same time, we strongly believe that human interpreters will remain vital, both as irreplaceable suppliers of relevant nuances and as models for automatic or partially automatic systems.
Human interpreting, too, will inevitably be carried out through the Web and its raising applications. Thus we foresee a continuing need for research on Web-based interpreting, and for data collection of realistic general-purpose or domain-oriented Web-based interpreting sessions.

2.2 Functionals of the ERIM human Interpreting platform

The ERIM-Interp network-based environment consists of a central communication server, two speaker stations, one interpreter station (cf. Fig. 1), with a multimodality server (exchange of short typed messages, whiteboard with shared pictures or files, and shared pointing and marking). To avoid complex problems due to turn overlap, we have adopted a push-to-talk discipline up to now.
The current implementation of ERIM-Interp, in Tcl/Tk, is platform independent (and runs on Windows, MacOS, eventually Linux), and uses an adapted version of the CommSwitch written by CMU for the CSTAR-II project.
It is also flexible: the CommServer can be hosted on a dedicated station or on any user workstation, two speakers may share the same station (in a "visit" situation), the scenario can be extended to include more than two interlocutors, more than one interpreter (in "one-way" interpreting situations), and hence possibly more than two languages.

3. Bilingual spontaneous speech collection

3.1 As the next step taken then, the ERIM Collecting platform

We have then developed the ERIM-Collect variant, intended to collect corpora (cf. Fig. 1), moreover to enhance collaborative generation and use of bilingual speech corpora; namely to:

- collect only "raw" data (web-based spontaneous dialogues in any language pairs), as multimodal as possible —with no built-in annotation scheme intended yet,
- motivate volunteers to produce the data,
- induce volunteering by offering free service (on one of the ERIM variants described here), in exchange for free data (users should agree to "donate their speech to science"),
- distribute the data as freeware (via GPL licensing) on the Web, in a "re-playable" form: for each dialogue, descriptors indicate essential (anonymous) facts about the participants, along with the list of turns, indications of files, speakers, and time stamps for each turn,
- make it possible for other researchers to enrich the corpora by adding annotations in parallel files, again sharable through the web; they might use an extended version of the "Replay" facility (cf. Fig. 3), with consensus on a shared file structure and XML descriptors format,
- develop the collection platform so that it can itself be offered as freeware on the Web.

Accordingly, ERIM-Collect (currently 350 Kbytes of code in Tcl/Tk) was defined as an extension of ERIM-Interp:

- ERIM-Collect is language-independent,
- data is recorded locally during the dialogue; speech files are in PCM 22kHz-16bit-mono format,
- session and speech turns descriptor files are now in XML format,
- after the conversation, local descriptors and files are transferred then structured in corpus bases on a Collection Server,
- everything possible should be recorded: speech, short texts, whiteboard events, video, objects which the speakers refer to (e.g. file names and urls). In the current version 3 of ERIM-Collect, voice and short texts are collected; whiteboard actions and video are currently added.

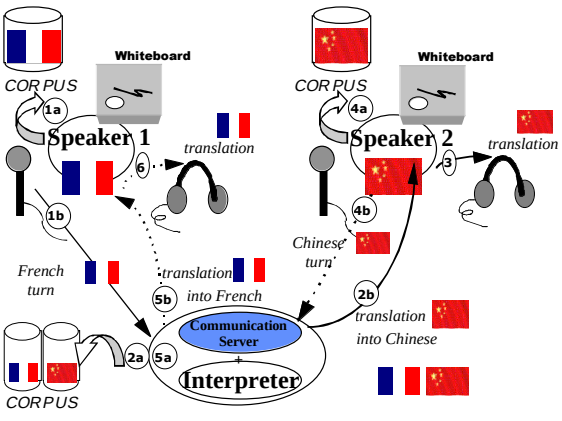


Figure 1: ERIM-Interp / ERIM-Collect

We describe here (cf. Fig. 1) a basic exchange within a French-Chinese collection session. First (1), the French interlocutor takes a turn of one or more utterances. This turn (speech, descriptors) is recorded locally (1a), and transmitted (1b) to the Interpreter and the CommServer which broadcasts it across the virtual room established for the conversation. The interpreter listens to the turn and (2) translates it into Chinese. The translated turn is recorded locally (2a) and broadcast (2b). The Chinese participant listens to the translation (3) and then answers (4). Again, his answer is stored locally and broadcast (4a and 4b). The interpreter then translates it into French (5) and the translation is stored locally (5a) and broadcast (5b).
In order to create various experimental settings, we may unlock the reception of some messages for some participants. For instance in (1b) the French voice could be made audible for the Chinese participant.

Figure 2 shows the screen which is presented to a conversational partner, as presently prototyped for the ERIM-Collect platform.

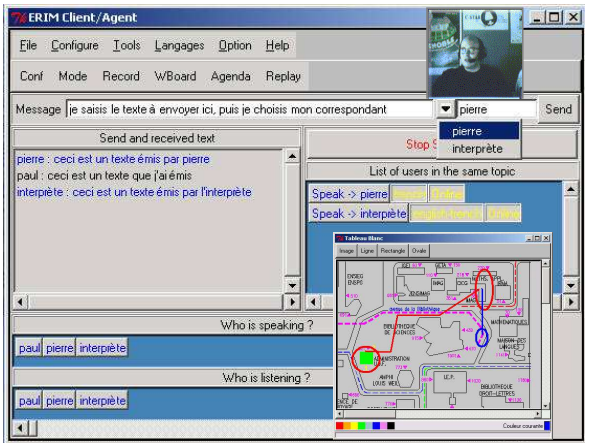


Figure 2: Speaker's screen

As for playback of a previously recorded bilingual dialogue, a full reconstruction is available. Simplified visual tracking is provided as shown in Figure 3. One can extract monolingual versions of the dialogues.
A first version of the DistribDial / Replay component (and web site) for such replays has just been completed.

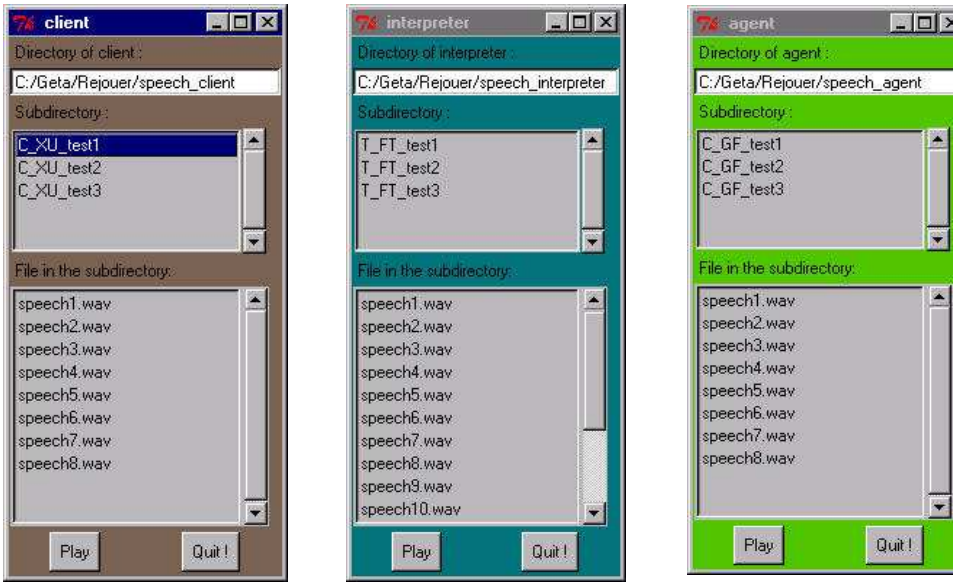


Figure 3: Playback of client, interpreter, and agent utterances

Successive versions of ERIM-Collect have been used for collecting first domain-oriented spontaneous speech corpora (hotel reservation) in Grenoble and Beijing (cf. 4.2).

3.2 Providing online aid to interpreters and/or speakers

In our "on demand interpretation" scenario, interpreters may be asked to jump from one conversation to another, and thus from one topic to another. This conversation switching is likely to be quite difficult, and stressful. Thus machine aids could be welcome: communication aids and language aids. We also envisage providing machine aids for the conversational partners, to help them do without interpreters so far as possible, if necessary.
The currently implemented "communication aids" include facilities to

- see and hear others (participants and interpreters),
- share data, possibly modifiable, markable, and "pointable" through the whiteboard,
- access an agenda for scheduling rendezvous.

Possible "language aids", to both the human interpreter and the speakers, are of three kinds:

- access to dictionaries via typed or voiced requests, and via automatic word spotting

followed by filtering, dictionary look-up, and presentation in a dedicated window,
- speech recognition, to alleviate difficulties of oral understanding when not using the interpreter, and to produce a log of the conversation (which can additionally help an interpreter jump in), after possible reduction,
- fully or partially automatic speech translation.

At this time most communication aids have been implemented. The scheduling agenda is global for an ERIM site, but each user handles it through a personalized view (cf. Fig. 4).

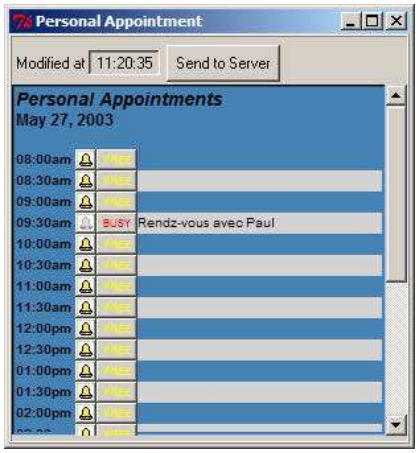


Figure 4: Window of user agenda

Language aids are the next step. An interface to existing free dictionary resources on the Papillon site [http://www.papillon-dictionary.org] should be added soon. A speech recognizer has been connected to the platform in another ERIM variant (the automatic interpretation pilot setup ERIM-paST). This Speech-To-Text facility could help as well to issue some draft transcripts during the dialogue.

3.3 Adding partially automatic Speech MT

An ERIM-paST (partially automated Speech Translation) platform is in progress at CLIPS in Grenoble, originally in cooperation with Spoken Translation Inc. (Berkeley). It aims at eventually providing some languge aids to speakers who "converse by themselves", and at allowing data recording of partially automatic interpreted dialogues (as a testing ground for Speech MT systems development, testing or tuning, at CLIPS). Experimentation with interactive disambiguation methods derived from the LIDIA project [Boitet & Blanchon, 1994] is also expected.
The detailed description of this ERIM variant is beyond the scope of this paper. Briefly stated, the goal is here a generic modular integration, through plug-in, of Speech MT modules (speech recognizers, text-to-text translators, speech synthesizers), either research components (for their fine testing and tuning) or off-the-shelf products. Objective is to carry out comparative assessment of their results, or possibly contrastive evaluation with the human production of an interpreter "warm body".
A first version of ERIM-paST is currently being prototyped, while integrating server-based (Philips, Linguatec, Scansoft) market components.

4. First corpus collection, towards a collaborative building/sharing scheme

4.1 Platform assessment: distant collection

Distant collection is also being tested, but in our first experiments Voice/IP still proved problematic when two turns overlapped. New efficient basic software and connection improvements are under evaluation. Record-then-send or record-while-sending (streaming) modes are available.
We may retain facilities for transmitting sound through phone lines. These might be used in operational contexts by telephone operators, such as Prosodie in France: since this company is also an Internet service provider, it can merge both tracks into a single communication.
Distant connection data is summarized in Figure 5.

Experiments (Grades from 0 to 5)	text	voice: record then send	voice: record & send (streaming)	voice: same with overlapping
Streaming	—	—	+	+
Connexion: Internet	100 Mbit	=	=	=
Reception quality	5	5	3	1
Speed of exchange	5	2	4	5
Reliability	5	5	4	1
Special problems / phenomena	None	User wary (too slow)	Some micro-cuts, but good overall quality	Unusable, bandwidth too large

Figure 5: Oral communication over the web

4.2 The ChinFaDial project, French-Chinese speech corpora

The system has been used in the ChinFaDial project for collecting bilingual French-Chinese interpreted spontaneous spoken dialogues, in the hotel reservation domain. This 3-year project was funded by LIAMA, a joint French-Chinese laboratory under both French INRIA and Chinese

CAS and MOST supervision. Our partner is the Chinese Information Processing group at NLPR (National Laboratory for Pattern Recognition), a research team within the Institute of Automation, Chinese Academy of Sciences (CAS-IA).

In ChinFaDial we have used intranets in Grenoble or in Beijing, with 3 participants using headsets, located in one or in 2 different buildings. It was possible for 2 speakers to share the same workstation, but we have mainly used the regular 3-station setting for the French-Chinese data collection. Some 10 hours of spontaneous translated spoken dialogues on "hotel information and reservation" have been recorded thus far. They produce about 43kBits per second.

Figure 5 shows a dialogue fragment transcription. We do not plan currently to transcribe or annotate corpora, but others will be very welcome to do so.

Participants to this first data collection have been at this time:

	Chinese	French	Total
Fr-Ch Interpreters	2	2	4
Interlocutors	3	3	6

There are 65 recorded dialogues with these characteristics:

	Minimum	Average	Maximum
Duration (sec)	457	635	874
Number of turns	28	52	78
Turn length (sec)	4	12	57

顾客/Client (7)
我在火车站：火车站离你们旅馆不知道远不远、怎么走？
(Je suis à la gare, je ne sais pas comment me rendre à l'hotel à partir de la gare.)

Agent/代理（7）
Alors c'est extrèmement simple,en sortant de la gare vous tournez à droite et c'est à 80 mètres en face de l'autre côté de la place.
（很简单、如果你出了火车站以后向右转、只要走到80米左右。就是我们的旅馆。）

顾客/Client (8)
好谢谢那就一会儿见
(Merci bien, alors à tout à l'heure)

Agent/代理（8）
Merci, bonsoir Monsieur,à tout à l'heure.
（谢谢…谢谢、那么、这是…一会见）

Figure 5: Dialogue between a French hotel manager and a Chinese client (manual transcript)

4.3 Ongoing developments, to promote collaborative corpus building

A website with a small 'DistribDial' server has been prototyped to freely distribute the sound files and their descriptors, and a Replay module. Our goal is to extend it to allow other groups to contribute to the site whatever annotations they may have created, and to share them under the same conditions (GPL). They should only agree to share a common file base structure and a flexible XML desciptor format for each annotation file.

Corpus collection in French-Chinese will extend. Further data collection using ERIM-Collect just started (spontaneous dialogues in French and Vietnamese, Tamil, Hindi), under support of AUF (University Agency for French-Speaking Communities), within the VTH-Fra.Dial project.

We are also considering distributing an ERIM-Collect "hardened" version on DistribDial, after strengthening robustness and usability, so that others can use it to do their own spoken dialogue collection.

4.4 Planned e-Training extensions: use of the platform to involve volunteer interpreters

Data collection being time-consuming all the same, our goal is not to do too much of it for its own sake, but to get it as byproduct of some "mutualized" use of the platform, in the open access mode.

Professional interpreters are unlikely to help on a non-profit basis, since interpreting is their livelihood. Improving junior interpreters or even advanced student interpreters, however, may find Web-based cooperation to be a good way of learning or perfecting their trade in real life situations.

We aim to induce volunteer interpreters or students of interpretation to translate bilingual dialogues online, by exchanging this on-line help for free use of our Web-based lab for *e*-training in interpretation.

We plan to develop an ERIM-Training variant platform, an *e*-training extension, with full recording of all speech interaction and any

multimodal event. Actually we already simulated the functional architecture of it, using the current ERIM-Collect in a multi-interpreter setting.
Different scenarios and settings can be envisaged. For example, in a distant training or practice situation, for a student interpreter: the student might be alone, gaining experience, or might be with an instructor, who could supervise or take over.
At the 2008 Olympic Games in Beijing, as another example, good student interpreters could be asked to aid bilingual communication in exchange for academic credit, and free tickets. Assume, for instance, that a French speaker and a Chinese speaker want to converse. They could then go to a PC, activate ERIM-Interp with ERIM-Assist for French-Chinese, click on the icon of an available interpreter, and begin a mediated conversation, which would be recorded if participants agree while using the service free of charge.

4.5 Building and sharing multilingual speech resources

We advocate and expect ERIM-Collect, once proposed in an open-access mode on the Web, to be willingly and freely operated by other researchers, under an agreed collaborative framework to be set up, with minimal method and technical consent on collecting procedures and corpus characteristic profiles, in order to bring building and sharing of raw multilingual speech corpora to a more rapid expansion.
Collaborative annotation work could take place as well, again with simple agreed procedures on content and descriptor files formats, and on a public use scheme.
Such tools, and their open use, could as well underlie valuable action towards supportive protection of "smaller languages", among others minor European languages, while for instance fostering distant learning of interpreting, and while easing the use of low-cost or even free interpreting facilities over the net.

5. Unification of ERIM platform variants

Work is now beginning on the integration of the different platforms presented here into one single multifunctional ERIMM system [Fafiotte & Boitet, 2003], for enhancing free multilingual multimodal network-based communication with distant interpreting and corpus collection.
Numerous technical issues arise in this effort. For instance, it is not immediately clear how the CommServer will accommodate server-based interactive lexical disambiguation during translation; or how to secure efficient streaming data transmission in a multicast scheme. Even so, the platform independence and plug-and-play generic architecture of ERIM set components make this integration effort quite realistic, in spite of the number and diversity of functions to be integrated.

Conclusion

We have presented several platforms developed in the long-range ERIM project. Each platform can aid in the study of spontaneous cross-lingual communication on the Web. The core platform is ERIM-Interp for Web-based human interpretation. ERIM-Collect is a deliberate development of the latter, dedicated to multilingual "raw" speech corpus building, and intended to alleviate the current scarcity of data —particularly open data—, and which can also support the construction of speech translation systems.
ERIM-Assist will add various machine aids for interpreters and conversational partners, while ERIM-paST (only briefly mentioned here) includes components for partially automatic speech translation.
We then reported on a first collection of spontaneous bilingual interpreted spoken dialogues for French-Chinese. This data, along with the collecting framework itself, will be distributed in the near future on the Web as shareware or GPL-ware, under a DistribDial component.
We are looking for funding to create ERIM-Training —a further extension of ERIM-Interp— which could serve as a valuable "Web-based language lab for interpreting" for distant *e*-training, while also providing new facilities for language learning.
We plan to continue research in the ERIM framework by collecting and distributing more data concerning more languages (Vietnamese, Tamil, Hindi to French). Data collection should be enhanced by a unified version of ERIM, offering all the functionalities of the platform variants.
More specifically, we hope that junior interpreters or advanced students in interpreting will volunteer to interpret and to practice with ERIM-Training, while users would agree to give their dialogues to science in exchange of using ERIM-Interp for free.

Acknowledgements

This work has been supported by CLIPS-IMAG (UJF University Grenoble 1, CNRS, INPG) and funded in part by the LIAMA French-Chinese Laboratory (ChinFaDial project), and by the Rhône-Alpes Region (ERIM project). Corpus collecting action is currently supported by AUF-LTT (University Agency for French-Speaking Communities, VTH-Fra.Dial project).
Our thanks go to Zhai JianShe (Nanjing

University, China) for early prototyping, to Julien Lamboley (at INSA, Lyon, France) for platform development, to members of the GETA and NLPR/CASIA-Beijing teams and to Brigitte Meillon at CLIPS-MultiCom, for their participation in data collection and related experiments.

References

Boitet C. & Blanchon H., 1994. *Multilingual Dialogue-Based MT for Monolingual Authors: the LIDIA Project and a First Mockup.* Machine Translation 9/2/94, pp. 99-132.

Coutaz J., Salber D., Carraux E. & Portolan N., 1996. *NEIMO, a Multiwork station Usability Lab for Observing and Analyzing Multimodal Interaction.* Proc. CHI'96 companion.

Fafiotte G. & Boitet C., 1994. *Report on first EMMI Experiments for the MIDDIM project in the context of Interpreting Telecommunications.* MIDDIM report TR-IT-0074 GETA-IMAG & ATR-ITL, Aug. 94, 11 p.

Fafiotte G. & Boitet C., 2003. *ERIMM, a platform for supporting and collecting multimodal spontaneous bilingual dialogues.* IEEE NLP-KE2003, Beijing, 26-29/10/03, 6 p.

Fafiotte G. & Zhai J.-S., 1999. *A Network-based Simulator for Speech Translation.* Proc. NPLRS'99, Beijing, 5-7/11/99, B. Yuan, T. Huang & X. Tang ed., pp. 511-514.

Furuse O., Sobashima Y., Takezama T. & Uratani N., 1994. *Bilingual corpus for speech translation.* Proc. AAAI-94 Workshop on Integration of Natural Language and Speech Processing, Seattle, Washington, USA, 31/7-1/8/94, ATR Interpreting Telecommunications.

Loken-Kim K.-H., Yato F. & Morimoto T., 1994. *A Simulation Environment for Multimodal Interpreting Telecommunications.* Proc. IPSJ-AV workshop, March 94, 5 p.

<url> C-STAR. http://www.c-star.org

<url> DARPA sites.
http://www.darpa.mil/ito/research/com/index.html, http://fofoca.mitre.org/doc.html

<url> GALAXY system architecture site. http://www.sls.lcs.mit.edu/sls/whatwedo/ architecture.html

<url> site web NESPOLE! http://nespole.itc.it

<url> site web PAPILLON.
http://www.papillon-dictionary.org

[<url> VERBMOBIL site. http://verbmobil.dfki.de

A Method of Creating New Bilingual Valency Entries using Alternations

Sanae Fujita **Francis Bond**
{sanae, bond}@cslab.kecl.ntt.co.jp
NTT Machine Translation Research Group
NTT Communication Science Laboratories
Nippon Telephone and Telegraph Corporation

Abstract

We present a method that uses alternation data to add new entries to an existing bilingual valency lexicon. If the existing lexicon has only one half of the alternation, then our method constructs the other half. The new entries have detailed information about argument structure and selectional restrictions. In this paper we focus on one class of alternations, but our method is applicable to any alternation. We were able to increase the coverage of the causative alternation to 98%, and the new entries gave an overall improvement in translation quality of 32%.

1 Introduction

Recently, deep linguistic processing, which aims to provide a useful semantic representation, has become the focus of more research, as parsing technologies improve in both speed and robustness (Uszkoreit, 2002). In particular, machine translation systems still mainly rely on large hand-crafted lexicons. The knowledge acquisition bottleneck, however, remains: precise grammars need information-rich lexicons, such as valency dictionaries, which are costly to build and extend. In this paper, we present a method of adding new entries to an existing bilingual valency dictionary, using information about verbal alternations.

The classic approach to acquiring lexical information is to build resources by hand. This produces useful resources but is expensive. This is still the approach taken by large projects such as FrameNet (Baker et al., 1998) or OntoSem. Therefore, there is a need to extend these handmade resources quickly and economically. Another approach is to attempt to learn information from corpora. There has been much research based on this, but due to the inevitable errors, there are few examples of lexicons being constructed fully automatically. Korhonen (2002) reports that the ceiling on the performance of mono-lingual subcategorization acquisition from corpora is generally around 80%, a level that still requires manual intervention. Yet another approach is to combine knowledge sources: for example to build a lexicon and then try to extend it using corpus data or to enrich monolingual data using multilingual lexicons (Fujita and Bond, 2002).

The aim of this research is not to create a lexicon from scratch, but rather to add further entries to an existing lexicon. We propose a method of acquiring detailed information about predicates, including argument structure, semantic restrictions on the arguments and translation equivalents. It combines two heterogeneous knowledge sources: an existing bilingual valency lexicon (the seed lexicon), and information about verbal alternations.

Most verbs have more than one possible argument structure (subcat). These can be regularized into pairs of alternations, where two argument structures link similar semantic roles into different subcats. Levin (1993) has identified over 80 alternation types for English, and these have been extended to cover 4,432 verbs in 492 classes (Dorr, 1997). In this paper, we will consider alternations between transitive (`Vt`) and intransitive (`Vi`) uses of verbs, where the subject of the intransitive verb (**S**) is the same as the object of the transitive verb (**O**) (e.g. *the acid dissolved the metal* ⇔ *the metal dissolved (in the acid)*) (Levin, 1993, 26–33)). We call the subject of the transitive verb **A** (ergative) and this alternation the **S**=**O** alternation.

Figure 1 shows a simplified example of an alternating pair in a bilingual valency dictionary (the valency lexicon from the Japanese-to-English machine translation system **ALT-J/E** (Ikehara et al., 1991)). This includes the subcategorization frame and selectional restrictions. As shown in Figure 1, Japanese, unlike English, typically morphologically marks the transitivity alternation.

We chose the **S**=**O** alternation because it is one

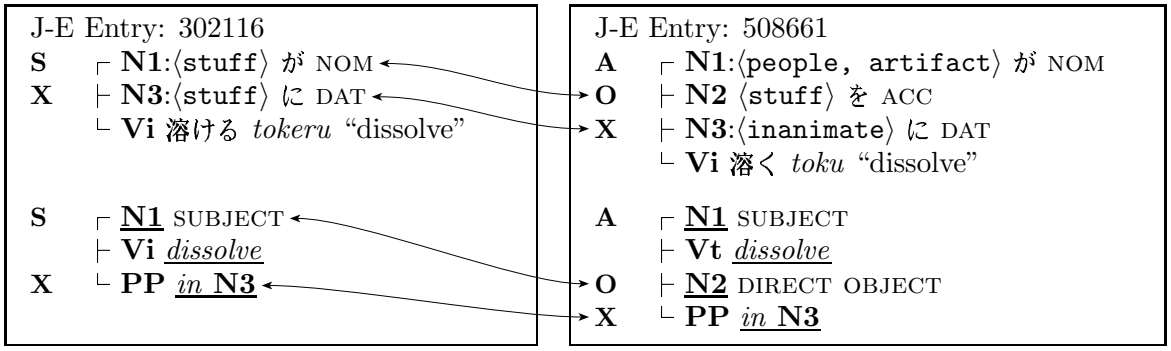


Figure 1: **Vi** 溶ける *tokeru* "dissolve" ↔ **Vt** 溶く *toku* "dissolve"

of the most common types of alternations, making up 34% of those discovered by Bond et al. (2002) and has been extensively studied. The method we present, however, can be used with any alternation for which lists of alternating verbs exist.

2 Resources

We use two main resources in this paper: (1) a seed lexicon of high quality hand-made valency entries; and (2) lists of verbs that undergo one or more **S**=**O** alternations.

The alternation list includes 449 native Japanese verbs that take the **S**=**O** alternation, based on data from Jacobsen (1981), Bullock (1999) and the Japanese/English dictionary EDICT (Breen, 1995). Each entry consists of a pair of Japanese verbs with one or more English glosses. Expanding out the English results in 839 Japanese-English pairs in all. Some examples are given in Table 1.

Intransitive			Transitive		
Ja		En	Ja		En
溶ける	tokeru	dissolve	溶く	toku	dissolve
泣く	naku	cry	泣かす	nakasu	make cry
上がる	agaru	rise	上げる	ageru	lift

Table 1: Verbs Undergoing the **S**=**O** Alternation

As a seed lexicon, we use the valency dictionary (Ikehara et al., 1997) from the Japanese-to-English machine translation system **ALT-J/E**. It consists of linked pairs of Japanese and English verbs. There are 5,062 Japanese verbs and 11,214 entries (ignoring all idiomatic and adjectival entries). Verb entries in both languages have information about the argument structure (subcat) of the verb. In addition to the core arguments, adjunct cases are added to many patterns to help in disambiguation.[1] The Japanese side has selectional restrictions (SR) on the arguments. The arguments are linked between the two languages using case-roles (**N1, N2, ...**).

The seed lexicon covered 381 out of the 449 linked Japanese pairs (85%). In the next section, in order to examine the nature of the alternation we compare the case roles and translation of the linked valency pairs.

3 The Nature of the S=O Alternation

3.1 Comparing Selectional Restrictions of A, O and S

In alternations, a given semantic role typically appears in two different syntactic positions: for example, the DISSOLVED role is the subject of intransitive *dissolve* and the object of the transitive. Baldwin et al. (1999) hypothesized that selectional restrictions (SRs) stay constant in the different syntactic positions. Dorr (1997), who generates both alternations from a single underlying representation, implicitly makes this assumption. In addition, Kilgarriff (1993) specifically makes the **A** ⟨`+sentient, +volition`⟩, while the **O** is ⟨`+changes-state, +causally affected`⟩.

However, we know of no quantitative studies of the similarities of alternating verbs. Exploiting the machine translation lexicon for linguistic research, we compare the SRs of **S** with both **A** and **O** for verbs that take the **S**=**O** alternation.

The SRs take the form of a list of semantic classes, strings or `*`. Strings only match specific words, while `*` matches anything, even non-nouns. The semantic classes are from the GoiTaikei ontology of 2,710 categories (Ikehara et al., 1997). It is an unbalanced hierarchy with a maximum depth of 12. The top node (level 1) is `noun`. The lower the level, the more specialized

[1] This is common in large NLP lexicons, such as COMLEX (Grishman et al., 1998). For example, the COMLEX 3.0 entry for *gather* notes that it coocurs with PPs headed by *around, inside, with, in* and *into*.

the meaning, and thus the more restrictive the SR.

We calculate the similarity between two SRs as the minimum distance (MD), measured as links in the ontology. If the SRs share at least one semantic class then the MD is zero. In this case, we further classified the SRs which are identical into "0 (Same)". For example, in Figure 1, the MD between **S** and **O** is "0 (Same)" because they have the same SR: ⟨`stuff`⟩. The MD between **A** and **S** is two because the shortest path from ⟨`artifact`⟩ to ⟨`stuff`⟩ traverses two links (`artifact` ⊂ `inanimate` ⊂ `stuff`).[2]

Figure 2 shows the MD between **O** and **S**, and **A** and **S**. The selectional restrictions are very similar for **O** and **S**. 30.1% have identical SRs, distance is zero for 27.5% and distance one is 28.3%. However, for **A** and **S**, the most common case is distance one (26.7%) and then distance two (21.5%). Although **O** and **S** are different syntactic roles, their SRs are very similar, reflecting the identity of the underlying semantic roles.

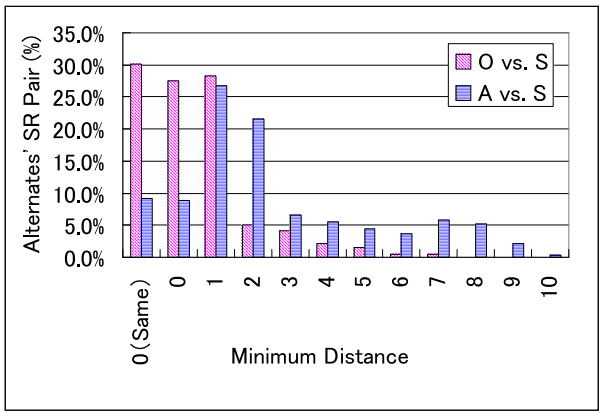


Figure 2: The Minimum Distance of Selectional Restrictions

Next, we examine whether **A**, **O**, and **S** are ⟨`+sentient, +volition`⟩ or not. In the GoiTaikei hierarchy, semantic classes subsumed by `agent` are ⟨`+sentient, +volition`⟩. **A** was very agentitive, with 60.1% of the SRs being subsumed by `agent`. The most frequent SR for **A** is ⟨`agent`⟩ itself (41.4%). **S** and **O** are less agentitive, with 13.9% and 14.1% of their respective selectional restrictions being agentitive. This data supports the hypothesis in Kilgarriff (1993).

[2]There is some variation due to lexicographer's inconsistencies. For example **X**'s SR is ⟨`stuff`⟩ in the intransitive and ⟨`inanimate`⟩ in the transitive entry. It should be ⟨`stuff`⟩ in both entries.

In summary, the SRs of **S** and **O** are not identical, but very similar. In comparison, **A** is more agentitive, and not closely linked to either.

3.2 Comparison of Japanese and English

From the point of view of constructing bilingual lexical entries, if the English main verb can translate both Japanese entries, then it is possible to automatically construct a usable English translation equivalent along with the Japanese alternation. In order to see how often this is the case, we compare Japanese and English alternations and investigate the English translations in the alternation list.

We divide the entries into five types in Table 2. The first three are those where the main English verb is the same. The most common type (30.0%) is made up of English unaccusative verbs which also undergo the **S**=**O** alternation [`S=O`]. The next most common (19.8%) is entries where the Japanese intransitive verb can be translated by making the transitive verb's English translation passive [`passive`]. In the third type (6.5%) the English is made transitive synthetically [`synthetic`]: a control verb (normally *make*) takes an intransitive verb or adjective as complement. The last two are those where either different translations are used (42.8%), or the same English verb is used but the valency change is not one of those described above.

The first three rows of Table 2 show the verbs whose alternate can be created automatically, 56.3% of the total. This figure is only an approximation, for two reasons. The first is that the translation may not be the best one, most verbs can have multiple translations, and we are only creating one. The second is that this upper limit is almost certainly too low. For many of the alternations, although our table contained different verbs, translations using identical verbs are also acceptable. In fact, most transitive verbs can be made passive, and most intransitive verbs embedded in a causative construction, so this alternative is always possible (and is also possible for Japanese). However, if the Japanese uses a lexical alternation, it is more faithful to link it to an English lexical alternation when possible.

4 Method of Creating Valency Entries

In this section we describe how we create new alternating entries. Given a verb, with dependents N_i, and an alternation that maps some or all of the N_i, we can create the alternate by analogy with existing alternating verbs. The basic flow of

Japanese		English Translation		English Structure		Type	No.	(%)
Vi	Vt	Vi	Vt	Vi	Vt			
弱まる	弱める	**S** weaken	**A** weaken **O**	***S*** *Vi*	***A*** *Vt* ***O***	S=O	138	30.0
漏れる	漏らす	**S** be omitted	**A** omit **O**	***S*** *be Vt-ed*	***A*** *Vt* ***O***	passive	91	19.8
泣く	泣かす	**S** cry	**A** make **O** cry	***S*** *Vi/be Adj*	***A*** *Vc* ***O*** *Vi/Adj*	synthetic	30	6.5
亡くなる	亡くす	**S** pass away	**A** lose **O**	***S*** *Vi*	***A*** *Vt* ***O***	Diff Head	197	42.8
じゃれる	じゃらす	**S** play	**A** play with **O**	***S*** *Vi*	***A*** *Vt prep* ***O***	Diff Struct	4	0.9

Vc is control verb such as *make,get,let,become*. Many entries also include information about non-core arguments/adjuncts.

Table 2: Classification of English Translations of the **S** = **O** Alternation List (Reference Data)

creating valency entries is as follows.

- For each dependent N_i

  if N_i participates in the alternation

  if N_i has an alternate in the target then map to it

  else delete N_i

  else transfer [non-alternating dependent]

- If the alternation requires a dependent not in the source
  Add the default argument

We use the most frequent argument in existing valency entries as a default. Specific examples of creating **S** = **O** alternations are given in the next section.

Although we only discuss the selectional restrictions and subcat information here, we also map the verb classes (given as verbal semantic attributes (Nakaiwa and Ikehara, 1997)). The mapping for the dependents in the alternation can be taken from existing lexical resources (Dorr, 1997), learned from corpora (McCarthy, 2000) or learned from existing lexicons (Bond et al., 2002).

4.1 Target

In this experiment, we look at one family of alternations, the **S** = **O** alternation. The candidate words are thus intransitive verbs with no transitive alternate, or transitive entries with no intransitive alternate. Alternations should be between senses, but the alternation list is only of words. Many of the candidate words (those that have a entry for only one alternate) have several entries. Only some of these are suitable as seeds. We don't use entries which are intransitive lemmas but have an accusative argument, which are intransitive (or transitive) lemmas but have an transitive translation (or intransitive), or which have both topic and nominative, such as (1), where the nominative argument is incorporated in the English translation.

(1)

N1:⟨animals⟩	は	N3:⟨"力"⟩	が
N1	*ha*	*N3:"chikara"*	*ga*
N1	TOP	N3:power	NOM

抜ける
nukeru
lose

N1 lose N1's energy

There are 115 entries (37 lemmas) which have only intransitive entries and 81 entries (25 lemmas) which have only transitive entries which are in our reference list of alternating verbs. We create intransitive entries using the existing transitive entries, and transitive entries using the existing intransitive entries.

4.2 Creating the Japanese subcat and SRs

In creating the intransitive entries from the transitive entries, we map the **O**'s SRs onto the **S**'s SRs, and change the case marker from accusative to nominative. We delete the **A** argument, and transfer any other dependents as they are.

In creating the transitive entries, we map the intransitive **S**'s SRs onto the new **O**'s SRs, and give it an accusative case-marker. If the intransitive entry has a demoted subject argument (where the Japanese case-marker is *ni* and the English preposition is *by*), we promote it to subject and use its SR for **A**. Otherwise we add a causative argument as ergative subject (**A**) with a default SR of ⟨agent⟩[3] and a nominative case-marker. We show an example in Figure 3.

4.3 Creating the English Equivalents

The English translation can be divided into three types: S=O, passive and synthetic. Therefore it is necessary to judge which type is appropriate for each entry, and then create the English. This judgement is shown in Figure 4. To judge whether an English

[3] ⟨agent⟩ is the most frequent SR for transitive verbs undergoing this alternation as shown in § 3.1.

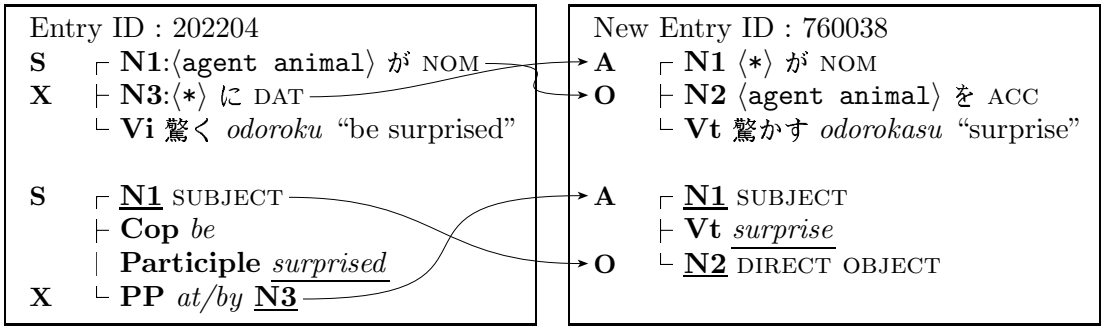


Figure 3: Seed: **Vi** 驚く *odoroku* "be surprised" ⇒ New entry: **Vt** 驚かす *odorokasu* "surprise"

verb could undergo the **S** = **O** alternation we used the LCS Database (EVCA+) (Dorr, 1997, `http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html`).

5 Evaluation

A total of 196 new entries were created for 62 verbs (25 `Vi` + 37 `Vt`) using the method outlined in § 4. We evaluated the quality by using the new entries in a machine translation system.

5.1 Translation-Based Evaluation

We evaluated the quality of the created entries in a translation-based regression test. We got two example sentences using each verb from Japanese newspapers and web pages: this gave a total of 124 test sentences. We translated the test sentences using **ALT-J/E**, both with (`with`) and without (`w/out`) the new entries.

Translations that were identical were marked `no change` (the system translates with a simple word dictionary if it has no valency entry). Translations that changed were evaluated by people fluent in both languages (two thirds by Japanese native speakers and one third by an English native speaker, not the authors). The translations were randomly presented to the evaluators labeled by `A` and `B`. Therefore evaluators did not know whether a translation is `with` or `w/out`. The translations were placed into three categories: (i) `A` is better than `B`, (ii) `A` and `B` are equivalent in quality, and (iii) `A` is worse than `B`. For example in (2), the evaluation was (iii). In this case `A` is `w/out` and `B` is `with`, so the new entry has improved the translation.

(2) 塩田 喜代子 さん は、 毛布 に
Shioda Kiyoko san wa, moufu ni
Shioda Kiyoko Ms. NOM blanket in
くるまり ながら。
kurumari nagara.
wrapped while.

(A) `Ms. Kiyoko Shioda is wrapped up to a blanket.`

(B) `Ms. Kiyoko Shioda is wrapped in a blanket.`

Table 3 shows the evaluation results, split into those for transitive and intransitive verbs. The most common result was that the new translation was `better` (46.0%). The quality was `equivalent` for 13.7% and `worse` for 14.5%. The overall improvement was 31.5% (46.0 − 14.5). Extending the dictionary to include the missing alternations gave a measurable improvement in translation quality.

	`Vi` Created		`Vt` Created		Total	
	No.	%	No.	%	No.	%
`better`	19	38.0	38	51.4	57	46.0
`equivalent`	5	10.0	12	16.2	17	13.7
`no change`	18	36.0	14	18.9	32	25.8
`worse`	8	16.0	10	13.5	18	14.5
Change		+22.0		+37.9		+31.5
Total	50	100.0	74	100.0	124	100.0

Table 3: Results of Translation-based Evaluation

5.2 Lexicographer's Evaluation

A manual analysis of a subset of the created entries was carried out by expert lexicographers familiar with the seed lexicon (not the authors). They found three major source of errors. The first was that alternation is a sense based phenomenon. As we built alternations for all patterns in the seed dictionary, this resulted in the creation of some spurious patterns. An example of an impossible entry is 捕らわれる *torawareru* "be caught", translated as *be picked up* with the inappropriate semantic restriction ⟨`concrete,material-phenomenon`⟩ on the subject. However, another good entry was cre-

Creating Intransitive entries:

if the original subcat has a control verb
(Vc ∈ {*make,have,get,cause*})

- **A** Vc **O** Vi/Adj
  ⇒ **S** Vi/be Adj `[synthetic]`
  (***A*** *make* ***O*** *cry* ⇒ ***S*** *cry*)

else (original head is Vt)

- if Vt undergoes the **S** = **O** alternation
  - **A** Vt **O** ⇒ **S** Vi `[S=O]`
    (***A*** *turn* ***O*** ⇒ ***S*** *turn*)
- else
  - **A** Vt **O** ⇒ **S** be Vt-ed `[passive]`
    (***A*** *injure* ***O*** *in* ***X***
    ⇒ ***S*** *be injured in* ***X***)

We made a special rule for the English Vt *have*. In this case the intransitive alternation will be *There is*: for example, 「及ぼす」 ***A*** *have* ***O*** *on X* ⇒ 「及ぶ」 *There be* ***S*** *on X*.

Creating Transitive Entries :
If the original subcat is:

- **S** Vi
  - if Vi undergoes the **S** = **O** alternation
    ⇒ **A** Vt **O** `[S=O]`
    (***S*** *spoil* ⇒ ***A*** *spoil* ***O***)
  - else ⇒ **A** Vc† **O** Vi `[synthetic]`
    (***S*** *rot* ⇒ ***A*** *make* ***O*** *rot*)
- **S** be Adj ⇒ **A** Vc† **O** Adj `[synthetic]`
  (***S*** *be prosperous* ⇒ ***A*** *make* ***O*** *prosperous*)
- **S** be Vt-ed ⇒ **A** Vt **O** (*by* **A**) `[passive]`
  (***S*** *be defeated (by* ***A****)* ⇒ ***A*** *defeat* ***O***)

† We use *make* as the control verb, Vc

Figure 4: Method of Creating English Side

ated, with the translation *be caught* and SRs ⟨`people,animal,artifact`⟩, and this was judged to be good.

The second source of errors was in the selectional restrictions. In around 10% of the entries, the lexicographers wanted to change the SRs. The most common change was to make the SR for **A** more specific than the default of `agent`.

The third source of errors was in the English translation, where the lexicographers sometimes preferred a different verb as a translation, rather than a regular alternation.

6 Discussion and Future Work

The above results show that alternations can be used to create rich and useful bilingual entries. In this section we discuss some of the reasons for errors, and suggest ways to improve and expand our method.

6.1 Rejecting Innappropriate Candidates

To make the construction fully automatic, a test for whether the Japanese side of the entry is appropriate or not is required.

One possibility is to add a corpus based filter: if no examples can be found that match the selectional restrictions for an entry, then it should be rejected. This could be done for each language individually. The problem with this approach is that many of the entries we created were for infrequent verbs. The average frequency in 16 years of Japanese newspaper text was only 173, and 22 verbs never appeared, although all were familiar to native speakers. We can, of course, use the web to alleviate the data sparseness problem.

6.2 Improving the English Translations

In this section we compare the distribution of the different types of translations for the reference data (§ 3.1) and the entries created by our method (§ 3.2). The breakdown is shown in Table 4. The first three rows show entries with the same English main verb.

One major discrepancy is in the frequency of the control verb construction. In `Vi`, no original transitive entry used control verbs. In general, when lexicographers create an entry, they prefer a simple entry to a synthetic one. Looking at the linguists' reference data, about 6.5% of the examples used control verbs. In the constructed data, 66.1% (77 entries) use the control verb *make*, more than any other category. For example, when the original intransitive entry is *N1 be exhausted*, *exhausted* is defined as adjective in the existing dictionary. So we create a new entry *N1 make N2 exhausted*$_{\text{adj}}$. However, there is a transitive verb *exhaust*, and it was preferred by the lexicographers: *N1 exhaust N2*. The algorithm needs to optionally convert adjectives to verbs in cases where there is overlap between the adjective and past participle.

Finally, we consider those Japanese alternations where the transitive and intransitive alternatives need translations with different English main verbs. A good example of this is `Vi` 亡くな

Type	English Structure `Vi`	`Vt`	Reference Data (Table2) No.	(%)	`Vi` Created No.	(%)	`Vt` Created No.	(%)
`S=O`	***S*** *Vi*	***A*** *Vt* ***O***	138	30.0	9	11.1	24	21.7
`passive`	***S*** *be Vt-ed*	***A*** *Vt* ***O***	91	19.8	71	87.7	14	12.2
`synthetic`	***S*** *Vi/be Adj*	***A*** *Vc* ***O*** *Vi/Adj*	30	6.5	0	0	76	66.1
Different Head			191	41.5	0	0.0	0	0.0
Different Structure			10	2.2	1	1.2	0	0.0
Total			460	100	81	100	115	100

Table 4: A Comparison of Reference Data with Created Alternations

る *nakunaru* "**S** pass away" and `Vt` 亡くす *nakusu* "**A** lose **O**".[4] These are impossible to generate using our method. Even with reliable English syntactic data, it would be hard to rule out *pass away* as a possible transitive verb or *lose* as an intransitive. They can only be ruled out by using data linking the subcat with the meaning, and this would need to be linked to the Japanese verbs' meanings. This may become possible with larger linked multi-lingual dictionaries, such as those under construction in the Papillon project,[5] but is not now within our reach.

In summary, we could improve the construction of the English translations by using richer English information, especially about past-participles or verb senses.

6.3 Usage as a Lexical/Translation Rule

Although we have investigated the use of alternations in lexicon construction, the algorithms could also be used directly, either as lexical/translation rules or to generate transitive and intransitive entries from a common underlying representation. For example, Shirai et al. (1999) uses the existing entries and lexical rules deploying them to translate causatives and passives (including adversative passives) from Japanese to English. Trujillo (1995) showed a method to apply lexical rules for word translation. That is, they expand the vocabulary using prepared lexical rules for each language, and create links for translation between the lexical rules of a pair of languages. Dorr (1997) and Baldwin et al. (1999) generate both alternates from a single underlying representation.

Our proposed method could partially be implemented as a lexical or a translation rule. But not all the word senses alternate (§ 4.2), and not all the target language entries are regularly translated by the same head (§ 3). Further many of the rules mix lexical and syntactic information, making them quite complicated. Because of that, it is easier to expand out the rules beforehand and enter them into the system.

6.4 Further Work

In this paper, we targeted native Japanese verbs only. **ALT-J/E** already has a very high coverage of native Japanese verbs. However, even in this case, we could increase the cover of this alternation from 85% to 98% (442 out of 449 alternation pairs now in the dictionary). Most valency dictionaries or new language pairs have less cover, and so will get more results. It is also possible to use this method so as to only create half the entries by hand, and then to automatically make the alternating halves (although not all the created entries will be perfect).

In addition to the native Japanese verbs, there are many Sino-Japanese verbal nouns that undergo **S**=**O** alternation (For example, (3) ↔ (4)).

(3) 店 が 製品 を 完売した
mise ga seihin o kanbai-shita
shop NOM products ACC sold out
The shop sold out of the products.

(4) 製品 が 完売した
seihin ga kanbai-shita
products NOM sold out
The products are sold out.

ALT-J/E's Japanese dictionary has about 2,400 verbal nouns which have usage as both transitive and intransitive. Of these only 536 are in the valency dictionary. Our next plan is to add them all to the valency dictionary, using alternations to make the process more efficient and consistent.

Another extension is to apply the method to other alternations, using either linguists' data or automatically acquired alternations (Oishi and Matsumoto, 1997; Furumaki and Tanaka, 2003;

[4] *My friend passed away* ↔ *I lost my friend.*
[5] `http://www.papillon-dictionary.org/`

McCarthy, 2000). In particular, **S** = **O** alternations make up only 34% of those discovered by Bond et al. (2002), we intend to investigate the alternations that make up the remainder.

7 Conclusion

We presented a method that uses alternation data to add new entries to an existing translation lexicon. The new entries have detailed information about argument structure and selectional restrictions. We were able to increase the coverage of the **S**=**O** alternation to 98%, and the new entries gave an overall improvement in translation quality of 32%.

References

Collin F. Baker, Charles J. Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics: COLING/ACL-98*, Montreal, Canada.

Timothy Baldwin, Francis Bond, and Ben Hutchinson. 1999. A valency dictionary architecture for machine translation. In *Eighth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99*, pages 207–217, Chester, UK.

Francis Bond, Timothy Baldwin, and Sanae Fujita. 2002. Detecting alternation instances in a valency dictionary. In *8th Annual Meeting of the Association for Natural Language Processing*, pages 519–522. The Association for Natural Language Processing.

Jim Breen. 1995. Building an electronic Japanese-English dictionary. Japanese Studies Association of Australia Conference (`http://www.csse.monash.edu.au/~jwb/jsaa_paper/hpaper.html`).

Ben Bullock. 1999. Alternative sci.lang.japan frequently asked questions. `http://www.csse.monash.edu.au/~jwb/afaq/jitadoushi.html`.

Bonnie J. Dorr. 1997. Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4):271–322.

Sanae Fujita and Francis Bond. 2002. A method of adding new entries to a valency dictionary by exploiting existing lexical resources. In *Ninth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2002*, pages 42–52, Keihanna, Japan.

Hisanori Furumaki and Hozumi Tanaka. 2003. The consideration of <n-suru> for construction of the dynamic lexicon. In *9th Annual Meeting of The Association for Natural Language Processing*, pages 298–301. (in Japanese).

Ralph Grishman, Catherine Macleod, and Adam Myers, 1998. *COMLEX Syntax Reference Manual*. Proteus Project, NYU. (`http://nlp.cs.nyu.edu/comlex/refman.ps`).

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–. In *Third Machine Translation Summit: MT Summit III*, pages 101–106, Washington DC. (`http://xxx.lanl.gov/abs/cmp-lg/9510008`).

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo. 5 volumes/CDROM.

Wesley Jacobsen. 1981. *Transitivity in the Japanese Verbal System*. Ph.D. thesis, University of Chicago. (Reproduced by the Indiana University Linguistics Club, 1982).

Adam Kilgarriff. 1993. Inheriting verb alternations. In *Sixth Conference of the European Chapter of the ACL (EACL-1993)*, pages 213–221, Utrecht. (`http://acl.ldc.upenn.edu/E/E93/E93-1026.pdf`).

Anna Korhonen. 2002. Semantically motivated subcategorization acquisition. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, Philadelphia, USA.

Beth Levin. 1993. *English Verb Classes and Alternations*. University of Chicago Press, Chicago, London.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the first Conference of the North American Chapter of the Association for Computational Linguistics. (NAACL)*, Seattle, WA.

Hiromi Nakaiwa and Satoru Ikehara. 1997. A system of verbal semantic attributes in japanese focused on syntactic correspondence between japanese and english. *Information Processing Society of Japan (IPSJ)*, 38(2):215–225. (In Japanese).

Akira Oishi and Yuji Matsumoto. 1997. Detecting the organization of semantic subclasses of Japanese verbs. *International Journal of Corpus Linguistics*, 2(1):65–89.

Satoshi Shirai, Francis Bond, Yayoi Nozawa, Tomiko Sasaki, and Hiromi Ueda. 1999. One method of fitting valency patterns to text. In *5th Annual Meeting of the Association for Natural Language Processing*, pages 80–83. The Association for Natural Language Processing.

Arturo Trujillo. 1995. Bi-lexical rules for multilexeme translation in lexicalist MT. In *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95*, pages 48–66, July.

Hans Uszkoreit. 2002. New chances for deep linguistic processing. In *19th International Conference on Computational Linguistics: COLING-2002*, pages XIV–XXVII, Taipei.

Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpora

Sylwia OZDOWSKA
Equipe de Recherche en Syntaxe et Sémantique
Université Toulouse le Mirail
5 allées Antonio Machado
31058 Toulouse Cedex 1 France
ozdowska@univ-tlse2.fr

Abstract

We present a word alignment procedure based on a syntactic dependency analysis of French/English parallel corpora called "alignment by syntactic propagation". Both corpora are analysed with a deep and robust parser. Starting with an anchor pair consisting of two words which are potential translations of one another within aligned sentences, the alignment link is propagated to the syntactically connected words. The method was tested on two corpora and achieved a precision of 94.3 and 93.1% as well as a recall of 58 and 56%, respectively for each corpus.

1 Introduction

It is now an aknowledged fact that parallel corpora, i.e. corpora made of texts in one language and their translation in another language, are well suited in particular to cope with the problem of the construction of bilingual resources such as bilingual lexicons or terminologies. Several works have focused on the alignment of units which are smaller than a sentence, for instance words or phrases, as to produce bilingual word, phrase or term associations. A common assumption is that the alignment of words or phrases raises a real challenge, since it is "neither one-to-one, nor sequential, nor compact", and thus "the correspondences are fuzzy and contextual" (Debili, 1997). Indeed, it is even often diffult for a human to determine which source unit correspond to which target unit within aligned sentences (Och and Ney, 2003).

Most alignment systems working on parallel corpora rely on statistical models, in particular the EM ones (Brown, Della Pietra and Mercer, 1993). Quite recently attempts have been made in order to incorporate different types of linguistic information sources into word and phrase alignment systems. The idea is to take into account the specific problems arising from the alignment at the word or phrase level mentioned in particular by Debili (1997). Different types of linguistic knowledge are exploited: morphological, lexical and syntactic ones. In the method described in this article, the syntactic information is the kernel of the alignment process. Indeed, syntactic relations identified on both sides of the French/English parallel corpus with a deep and robust parser are used to find out new correspondences between words or to confirm existing ones in order to achieve a high accuracy alignment. We call this procedure "alignment by syntactic propagation".

2 State of the art

2.1 Term alignment

Two kinds of methods have been basically proposed in order to address the problem of bilingual lexicon extraction. On the one hand, terms are recognized in both source and target language and then they are mapped to each other (Daille, Gaussier and Langé, 1994). On the other hand, only source terms are extracted and the target ones are discovered through the alignment process (Gaussier, 1998; Hull, 2001). The alignment between terms is obtained either by computing association probabilities (Gaussier, 1998 ; Daille, Gaussier and Langé, 1994) or by identifying, for a given source term, a sequence of words in the target language which is likely to contain or to correspond to its translation (Hull, 2001). In so far as the precision rate may be affected by the number of alignments obtained (Daille, Gaussier and Langé, 1994; Gaussier, 1998), the results achieved basically range between 80% and 90%, for the first 500 alignments. As for the method described in (Hull, 2001), the precison reported is 56%.

It should be noticed that the use of linguistic knowledge is most of the time restricted to the term recognition stage. This kind of knowledge is quite rarely taken into account within the very alignment process, except for the approach implemented by Daille, Gaussier and Langé (1994), which try to take advantage of

correspondences between the syntactic patterns defined for each language.

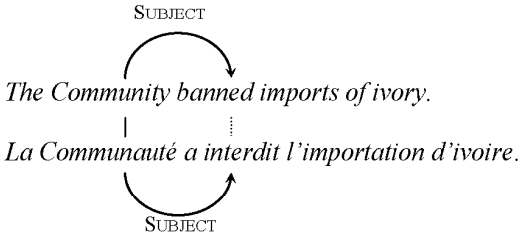
2.2 Word alignment

Quite recently attempts have been made in order to incorporate different types of linguistic information sources into word alignment systems and to combine them with statistical knowledge. Various and more or less complex sources of linguistic knowledge are exploited: morphological, lexical (Arhenberg, Andersson and Merkel, 2000) and syntactic knowledge (Wu, 2000; Lin and Cherry, 2003). The contribution of these information sources to the alignment process with respect to the statistical data varies according to the considered system. However, as pointed out by Arhenberg, Andersson and Merkel (2000) as well as Lin and Cherry (2003), the introduction of linguistic knowledge leads to a significant improvement in alignment quality. In the first case, the accuracy goes from 91% for a baseline configuration up to 96.7% for a linguistic knowledge based one. In the second, the precision rate is increased from 82.7% up to 89.2% and the improvement noticed have been confirmed within the framework of an evaluation task (Mihalcea and Pedersen, 20003).

For our part, we propose a method in which the syntactic information plays a major role in the alignment process, since syntactic relations are used to find out new correspondences between words or to confirm the existent ones. We chose this approach in order to achieve a high accuracy alignment both at word and phrase level. Indeed, we aim at capturing frequent alignments between words and phrases as well as those involving sparse or corpus specific ones. Moreover, as stressed in previous works, using syntactic dependencies seems to be particularly well suited to solve *n*-to-1 or *n*-to-*m* alignments (Fluhr, Bisson and Elkateb, 2000) and to cope with the problem of linguistic variation and non correspondence across languages, for instance when aligning terms (Gaussier, 2001).

3 Starting hypothesis

We take as a starting point the hypothesis formulated by Debili and Zribi (1996) according to which "*paradigmatic connections can help to determine syntagmatic relations, and conversely*"[1]. More precisely, the idea is that one can make use of syntactic relations to validate or invalidate the existence of alignment links, on the one hand, and to create new ones, on the other hand. The reasoning is as follows : if there is a pair of anchor words, i.e. if two words $w1_i$ (*community* in the example) and $w2_m$ (*communauté*) are aligned at the sentence level, and if there is a syntactic relation standing between $w1_i$ (*community*) and $w1_j$ (*ban*) on the one hand, and between $w2_m$ (*communauté*) and $w2_n$ (*interdire*) on the other hand, then the alignment link is propagated from the anchor pair (*community, communauté*) to the words (*ban, interdire*). We call this procedure "alignment by syntactic propagation".



In the rest of this article, we describe the overall design and implementation of the syntactic propagation process and the results of applying it to two parsed French/English parallel corpora: INRA and JOC.

4 Corpus processing

The alignment by syntactic propagation was tested on two different parallel corpora aligned at the sentence level: INRA and JOC. The first corpus was constituted at the National Institute for Agricultural Research (INRA)[2] to enrich a bilingual terminology database exploited by translators. It comprises about 300,000 words and mainly consists of research and popular-science papers, press releases.

The JOC corpus was provided by the ARCADE project, a campaign devoted to the evaluation of parallel text alignment systems (Veronis and Langlais, 2000). It contains written questions on a wide variety of topics addressed by members of the European Parliament to the European Commission and corresponding answers published by the Official Journal of the European Community in nine official languages. A portion of about 400,000 words of the French and English parts were used in the framework of the ARCADE evaluation task.

The corpus processing was carried out by a French/English parser: SYNTEX (Bourigault and Fabre, 2000; Frérot, Fabre and Bourigault, 2003). SYNTEX is a dependency parser whose input is a

[1]Our translation of the French version *« les liaisons paradigmatiques peuvent aider à déterminer les relations syntagmatiques, et inversement »*.

[2] We are grateful to A. Lacombe who allowed us to use this corpus for research purposes.

POS tagged[3] corpus–meaning each word in the corpus is assigned a lemma and grammatical tag. The parser identifies syntactic dependencies in the sentences of a given corpus, for instance subjects, direct and indirect objects of verbs. Once all syntactic dependencies have been identified, a set of words and phrases is extracted out of the corpus.

Both versions of the parser–the French one and the English one–are being developed according to the same procedures and architecture. The parsing is performed independently in each language, yet the outputs are quite homogeneous since the syntactic dependencies are identified and represented in the same way in both languages. In this respect, the alignment method proposed is different from the ones developed by Wu (2000) as well as Lin and Cherry (2003): the former is based on synchronous parsing while the letter uses a dependency tree generated only in the source language.

In addition to parsed French/English corpus aligned at the sentence level, the syntactic alignment requires pairs of anchor words be identified prior to propagation as to start the process. In this study, we chose to extract a lexicon out of the corpus, the anchor pairs being located both by projecting the lexicon at the level of aligned sentences and processing the identical and fuzzy cognates.

5 Identification of anchor pairs

To derive a list of words which are likely to be used to initiate the syntactic propagation process out of the corpus, we implemented a widely used method described notably in (Gale and Church, 1991; Ahrenberg, Andersson and Merkel, 2000) which is based on the assumption that the words which appear frequently in aligned text segments are potential translation equivalents. For each source (English) and target (French) unit, respectively u_1 and u_2, extracted by SYNTEX, the translation equivalents are searched for by counting co-occurrences of (u_1, u_2) in aligned sentences in comparison with their overall occurrences in the corpus and then an association score is computed. In this study, we chose the Jaccard association score which is calculated as follows:

$$j(u_1, u_2) = \frac{f(u_1, u_2)}{f(u_1) + f(u_2) - f(u_1, u_2)}$$

[3] We use both the French and English versions of the Treetagger. (http://www.ims.uni-stuttgart.de)

The association score is computed provided the number of overall occurrences of u1 and u2 is higher than 4 since statistical techniques have proved to be particularly efficient when aligning frequent units. Moreover, the alignments are filtered according to the $j(u_1, u_2)$ value, provided the latter is higher than 0.2. Then, two tests, based on cognate recognition and mutual correspondence condition (Altenberg, 1999), are applied as to filter spurious associations out of the initial lexicon.

The identification of anchor pairs, consisting of words which are translation equivalents within aligned sentences, combines both the projection of the initial lexicon and the recognition of cognates for words which have not been taken into account in the lexicon. These pairs are used as the starting point of the propagation process.

Table 1 gives some characteristics of the two corpora as for the number of aligned sentences, the overall number of anchor pairs identified, the average number of anchor pairs per sentence pair as well as the precision rate[4] of the anchor pairs. It can be seen that a high number of anchor pairs has been identified per sentence for both corpora with a high accuracy.

	INRA	JOC
aligned sentences	7056	8774
anchor pairs	42570	58771
words/source sentence	21	25
words/target sentence	24	30
anchor pairs/sentence	6.38	6.77
precision (%)	98	99.3

Table 1: The identification of anchor pairs

6 Syntactic propagation

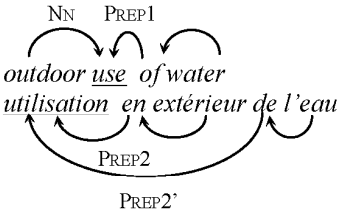
6.1 Two types of propagation

The syntactic propagation may be performed according to two different directions. Indeed, a given word is likely to be both governor and dependent with respect to other words. The first direction consists in starting with dependent anchor words and propagating the alignment link to the governors (DepGov propagation). The DepGov propagation is *a priori* not ambiguous since one dependent is governed at most by one word. Thus, there is just one syntactic relation on which the propagation can be based. The syntactic structures are said to be parallel in English and French provided the two following conditions are met: i) the relation under consideration is identical in both languages and ii) the words involved in the

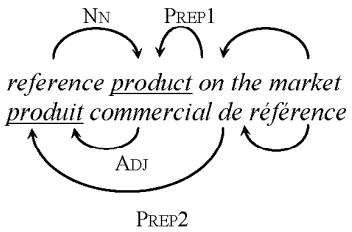
[4] The precision was evaluated manually

syntactic propagation have the same POS. The second direction goes the opposite way: starting with governor anchor words, the alignment link is propagated to the dependents (GovDep propagation). In this case, several relations which may be used to achieve the propagation are available, as it is possible for a governor to have more than one dependent, and so the propagation is potentially ambiguous. The ambiguity is particularly widespread when performing the GovDep propagation from head nouns to their nominal and adjectival dependents. Let us consider the example (1). There is one occurrence of the relation PREP in English and two in French. Thus, it is not possible to determine *a priori* whether to propagate using the relations NN/PREP2, on the one hand, and PREP1/PREP2', on the other hand, or NN/PREP2' and PREP1/PREP2. Moreover, even if there is just one occurrence of the same relation in each language, it does not mean that the propagation is of necessity performed through the same relation, as shown in example (2).

(1)



(2)



In the following sections, we describe precisely the implementation of the two types of propagation defined above in order to align verbs (section 6.2), on the one hand, and nouns and adjectives, on the other hand (section 6.3). To this, we rely on different propagation patterns. Propagation patterns are given in the form CDep-REL-CGov, where CDep is the POS of the dependent, REL is the syntactic relation and CGov, the POS of the governor. The anchor element is underlined and the one aligned by propagation is bolded. For instance, the pattern N-SUJ-**V** corresponds to the propagation going from a noun anchor pair to the verbs through the subject relation.

6.2 Alignment of verbs

Verbs are aligned according to eight propagation patterns, that is to say five for the DepGov propagation and three for the GovDep one.

DEPGOV PROPAGATION TO ALIGN GOVERNOR VERBS. Five propagation patterns are used to align verbs: Adv-MOD-**V** (1), N-SUJ-**V** (2), N-OBJ-**V** (3), N-PREP-**V** (4) and V-PREP-**V** (5).

(1) The net ***is*** *then* ***hauled*** *to the shore.*
Le filet ***est*** *ensuite* ***halé*** *à terre.*
(2) The fish ***are*** *generally* ***caught*** *when they migrate from their feeding areas.*
Généralement les poissons ***sont capturés*** *quand ils migrent de leur zone d'engraissement.*
(3) Most of the young shad ***reach*** *the sea.*
La plupart des alosons ***gagne*** *la mer.*
(4) The eggs are very small and ***fall*** *to the bottom.*
Les oeufs de très petite taille ***tombent*** *sur le fond.*
(5) X is a model which ***was designated*** *to stimulate...*
X est un modèle qui ***a été conçu*** *pour stimuler...*

GOVDEP PROPAGATION TO ALIGN DEPENDENT VERBS. The alignment links are propagated from the dependents to the verbs using three propagation patterns: **V**-PREP-V (1), **V**-PREP-N (2) and **V**-PREP-Adj (3).

(1) Ploughing tends to ***destroy*** *the soil microaggregated structure.*
Le labour tend à ***rompre*** *leur structure microagrégée.*
(2) The capacity to ***colonize*** *the digestive mucosa...*
L'aptitude à ***coloniser*** *le tube digestif...*
(3) An established infection is impossible to ***control****.*
Toute infection en cours est impossible à ***maîtriser****.*

	DepGov propagation	GovDep propagation
INRA		
precision (%)	94.1	96.7
JOC		
precision (%)	92.7	97.5

Table 2: Alignment of verbs by means of the DepGov and GovDep propagation

6.3 Alignment of adjectives and nouns

As for verbs, the two types of propagation described in section 6.1 are used to align adjectives and nouns. However, as far as these categories of words are concerned, they can't be treated in a

fully independent way when propagating from head noun anchor words in order to align the dependents. Indeed, the syntactic structure of noun phrases may be different in English and French, since they rely on a different type of composition to produce compounds and on the same one to produce free noun phrases (Chuquet and Paillard, 1989). Then the potential ambiguity arising from the GovDep propagation from head nouns evoked in section 6.1 may be accompanied by variation phenomena affecting the category of the dependents, called transposition (Vinay and Darbelnet, 1958; Chuquet and Paillard, 1989). For instance, a noun may be rendered by an adjective, or vice versa: *tax treatment profits* is translated by *traitement fiscal des bénéfices*, so the noun *tax* is in correspondence with the adjective *fiscal*. The syntactic relations used to propagate the alignment links are thus different.

In order to cope with the variation problem, the propagation is performed whether the syntactic relations are identical in both languages or not, and if they are not, whether the categories of the words to be aligned are the same or not. To sum up, adjectives and nouns are aligned *separately* of each other by means of DepGov propagation or GovDep propagation provided that the governor is not a noun. They are *not* treated *separately* when aligning by means of GovDep propagation from head noun anchor pairs.

DEPGOV PROPAGATION TO ALIGN ADJECTIVES. The propagation patterns involved are: Adv-MOD-**Adj** (1), N-PREP-**Adj** (2) and V-PREP-**Adj** (3).

(1) The white cedar exhibits a very ***common*** *physical defect.*
Le Poirier-pays présente un défaut de forme très ***fréquent****.*
(2) The area presently ***devoted*** *to agriculture represents...*
La surface actuellement ***consacrée*** *à l'agriculture représenterait...*
(3) Only fours plots were ***liable*** *to receive this input.*
Seulement quatre parcelles sont ***susceptibles*** *de recevoir ces apports.*

DEPGOV PROPAGATION TO ALIGN NOUNS. Nouns are aligned according to the following propagation patterns: Adj-ADJ-**N** (1), N-NN-**N**/N-PREP-**N** (2), N-PREP-**N** (3) and V-PREP-**N** (4).

(1) Allis shad remain on the continental ***shelf****.*
La grande alose reste sur le ***plateau*** *continental.*
(2) Nature of micropolluant ***carriers****.*
La nature des ***transporteurs*** *des micropolluants.*
(3) The ***bodies*** *of shad are generally fusiform.*
Le ***corps*** *des aloses est généralement fusiforme.*
(4) ***Ability*** *to react to light.*
Capacité *à réagir à la lumière.*

	DepGov propagation	
	Adjectives	Nouns
INRA		
precision (%)	98.7	94.2
JOC		
precision (%)	97.2	93.7

Table 3: Alignment of adjectives and nouns by means of the DepGov propagation

UNAMBIUOUS GOVDEP PROPAGATION TO ALIGN NOUNS. The propagation is not ambiguous when dependent nouns are not governed by a noun. This is the case when considering the following three propagation patterns: **N**-SUJ|OBJ-V (1), **N**-PREP-V (2) and **N**-PREP-Adj (3).

(1) The ***caterpillars*** *can inoculate the* ***fungus****.*
Les ***chenilles*** *peuvent inoculer le* ***champignon****.*
(2) The roots are placed in ***tanks****.*
Les racines sont placées en ***bacs****.*
(3) Botrysis, a fungus responsible for grey ***rot****.*
Botrysis, champignon responsable de la ***pourriture*** *grise.*

POTENTIALLY AMBIGUOUS GOVDEP PROPAGATION TO ALIGN NOUNS AND ADJECTIVES. As we already explained in section 6.1, the propagation is potentially ambiguous when starting with head noun anchor words and trying to align the noun(s) and/or adjective(s) they govern. Considering this potential ambiguity, the algorithm which supports GovDep propagation form head noun anchor words (*n1, n2*) takes into account three situations which are likely to occur :

1. if each of *n1* and *n2* have only one dependent, respectively *reg1* and *reg2,* involving one of the following relations NN, ADJ or PREP; *reg1* and *reg2* are aligned;

   the drained whey
   le lactosérum d'égouttage
   ⇒ (***drained, égouttage***)

2. *n1* has one dependent *reg1* and *n2* several ones $\{reg2_1, reg2_2, \ldots, reg2_n\}$, or vice versa. For each $reg2_i$, check if one of the possible alignments has already been

performed, either by propagation or anchor word spotting. If such an alignment exists, remove the others ($reg1$, $reg2_k$) such as $k \neq i$, or vice versa. Otherwise, retain all the alignments ($reg1$, $reg2_i$), or vice versa, without solving the ambiguity;

stimulant substances which are absent from...
substances solubles stimulantes absentes de...
(*stimulant*, {*soluble, stimulant, absent*})
already_aligned(*stimulant, stimulant*) = 1
⇒ ***stimulant, stimulant***)

3. both *n1* and *n2* have several dependents, {$reg1_1$, $reg1_2$, ..., $reg1_m$} and {$reg2_1$, $reg2_2$, ..., $reg2_n$} respectively. For each $reg1_i$ and each $reg2_j$, check if one/several alignments have already been performed. If such alignments exist, remove all the alignments ($reg1_k$, $reg2_l$) such as $k \neq i$ or $l \neq j$. Otherwise, retain all the alignments ($reg1_i$, $reg2_j$) without solving the ambiguity.

unfair trading practices
pratiques commerciales déloyales
(*unfair*, {*commercial, déloyal*})
(*trading*, {*commercial, déloyal*})
already_aligned(*unfair, déloyal*) = 1
⇒ (***unfair, déloyal***)
⇒ (***trading, commercial***)

a big rectangular net, which is lowered...
un vaste filet rectangulaire immergé...
(*big*, {*vaste, rectangulaire, immergé*})
(*rectangular*, {*vaste, rectangulaire, immergé*})
already_aligned(*rectangular, rectangulaire*) = 1
⇒ (***rectangular, rectangulaire***)
⇒ (*big*, {*vaste, immergé*})

The implemented propagation algorithm has two major advantages: it allows to solve some alignment ambiguities taking advantage of alignments which have been performed previously. This algorithm allows also to cope with the problem of non correspondence between English and French syntactic structures and makes it possible to align words using different syntactic relations in both languages, even though the category of the words under consideration is different.

	GovDep propagation	
	Gov≠Noun	Gov=Noun
INRA		
precision (%)	95.4	97.7
JOC		
precision (%)	95	95.4

Table 4: Alignment of adjectives and nouns by means of the GovDep propagation

6.4 Overall results

Table 5 gives a summary of the results obtained by applying all propagation patterns according to each corpus. It can be seen that the highest accuracy is achieved for the alignments corresponding to anchor pairs validated by the syntactic propagation (AP and PP): 99.7 and 99.8% precision, respectively for INRA and JOC. The rates tend to decrease – respectively 88.5 and 86.1% – as regards alignments established only by means of propagation, referred to as propagated pairs (PP), and is even lower – 76.3% – for the anchor pairs which have not been confirmed by the propagation (AP). Furthermore, the new alignments produced account for less than 20% of overall alignments to approximately 50% for the confirmed ones. Finally, since the method aims at aligning content words, the recall is assessed in relation to their overall occurrences in the corpora.

	Total	AP	AP and PP	PP
INRA				
alignments	50438 (100%)	23646 (47%)	18923 (37%)	7868 (16%)
precision (%)	94.3	76.3	99.7	88.5
recall (%)	58			
JOC				
alignments	71814 (100%)	37118 (52%)	21625 (30%)	13073 (18%)
precision (%)	93.1	94	99.8	86.1
recall (%)	56			

Table 5: overall results of word alignment

7 Discussion

The results achieved by the syntactic propagation method are quite encouraging. They show a high global precision rate – 94.3% for the INRA corpus and 93.1% for the JOC – assessed respectively against a reference list of approximately 8000 and 4600 alignments.

Various reasons make it difficult to compare the results of this experiment with those reported in the literature and presented in section 2. Indeed, each approach has been tested on a different corpus and the results achieved could depend on the type of texts used for evaluation purposes. Moreover, the reference alignment lists, i.e. the gold standards, have probably been established according to different annotation criteria, which could also influence the quality of the results. Finally, each system has been designed, or at least used, to perform a specific task and evaluated in this respect. Daille, Gaussier and Langé (1994), as well as Gaussier (1998) and Hull (2001), were interested in bilingual terminology extraction so that word alignment could not be considered as an end in itself but rather as a basis for term alignment. The system proposed by Wu (2000) aims at bilingual language modelling, word and phrase alignment is incorporated as a subtask. Finally, Arhenberg, Andersson and Merkel (2000) as well as Lin and Cherry (2003) addressed the problem of full word alignment without restricting themselves to content words. Both noticed that the integration of linguistic knowledge, morphological and lexical for the former, syntactic for the latter, improves the alignment quality. However, concerning the approach proposed by Lin and Cherry (2003), it should be pointed out that linguistic knowledge is considered secondary to statistical information. As regards the alignment by syntactic propagation, linguistic knowledge is the kernel of the approach rather than an additional information.

The propagation of alignments links using syntactic relations has proved very efficient when the same propagation pattern is used in both languages, i.e. when the syntactic structures are identical. A high level of precision is also achieved in the case of noun/adjective transpositions, even if the category of the words to be aligned varies. We are actually pursuing the study of non-correspondence between syntactic structures in English and French outlined in (Ozdowska and Bourigault, 2004). The aim is to determine whether there are some regularities in rendering certain English structures into certain French ones or not. If variation across languages is subjected to such regularities, the syntactic propagation could then be extended to the cases of non correspondence.

References

Ahrenberg L., Andersson M. and Merkel M. 2000. A knowledge-lite approach to word alignment, Véronis J. (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 97-138.

Altenberg B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences, Hasselgard and Oksefjell (eds), pp. 249-268.

Bourigault D. and Fabre C. 2000. Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, pp. 131-151, Université Toulouse le Mirail.

Brown P., Della Pietra S. and Mercer R. 1993. *The mathematics of statistical machine translation : parameter estimation*, Computational Linguistics, 19(2), pp. 263-311.

Chuquet H. and Paillard M. 1989. *Approche linguistique des problèmes de traduction anglais/français*, Ophrys.

Daille B., Gaussier E. and Langé J.-M. 1994. Towards Automatic Extraction of Monolingual and Bilingual Terminology, *Proceedings of the International Conference on Computational Linguistics* (COLING'94) , pp. 515-521.

Debili F. 1997. L'appariement :quels problèmes ?, *Actes des 1ères JST 1997 FRANCIL de l'AUPELF-UREF*, pp. 199-206.

Debili F. and Zribi A. 1996. Les dépendances syntaxiques au service de l'appariement des mots, *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle* (RFIA'96).

Fluhr C., Bisson B. and Elkateb F. 2000. Parallel Text Alignment Using Crosslingual Information Retrieval Techniques, Véronis, J. (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers.

Fox H. J. 2002. Phrasal Cohesion and Statistical Machine Translation, *Proceedings of EMNLP-02*, pp. 304-311.

Frérot C., Bourigault D. and Fabre C. 2003. Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus. Le cas du rattachement verbal à distance de la préposition « de », in *Traitement Automatique des Langues*, 44-3.

Frérot C., Rigou G. and Lacombe A. 2001. Approche phraséologique d'une extraction automatique de terminologie dans un corpus scientifique bilingue aligné, *Actes des 4èmes rencontres Terminologie et Intelligence Artificielle*, Nancy, pp 180-188.

Gale W. A. and Church K. W. 1991. Identifying Word Correspondences in Parallel Text,

Proceedings of the DARPA Workshop on Speech and Natural Language.

Gaussier E. 1998. Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora, *Proceedings of the joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics* (COLING/ACL'98) , pp. 444-450.

Gaussier E. 2001. General considerations on bilingual terminology extraction, D. Bourigault, Ch. Jacquemin, M.–C. L'Homme (Eds.), *Recent Advances in Computational Terminology*, John Benjamins, pp. 167-183.

Harris B. 1988. Bi-text, A New Concept in Translation Theory, *Language Monthly*, 54, pp.8-10.

Hull D. 2001. Software tools to support the construction of bilingual terminology lexicons, Bourigault, D., Jacquemin, Ch. and L'Homme, M.-C. (Eds.), *Recent Advances in Computational Terminology*, John Benjamins, pp. 225-244.

Lin D. and Cherry C. 2003a. Linguistic Heuristics in Word Alignment, *Proceedings of PACLing 2003*.

Lin D. and Cherry C. 2003b. ProAlign: Shared Task System Description, *Workshop Proceedings on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond (HLT-NAACL 2003)*.

Mihalcea R. and Pedersen T. 2003. An Evaluation Exercise for Word Alignment, *Workshop Proceedings on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond (HLT-NAACL 2003)*, pp. 1-10

Och F. Z. and Ney H., 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1), pp. 19-51.

Ozdowska S. and Bourigault D. 2004. Détection de relations d'appariement bilingue entre termes à partir d'une analyse syntaxique de corpus, *Actes du 14ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence artificielle*

Véronis J. (Ed). 2000. *Parallel Text Processing : Alignment and Use of Parallel Corpora*, Dordrecht : Kluwer Academic Publishers.

Véronis J. and Langlais P. 2000. Evaluation of parallel text alignment systems. The ARCADE project, Véronis J. (ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 371-388

Vinay J-P. and Darbelnet J. 1958. *Stylistique comparée du français et de l'anglais* , Paris, Didier.

Wu D. 2000. Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars, Véronis, J. (Ed.), *Parallel Text Processing : Alignment and Use of Translation Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 139-167.

Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications

Kiyotaka Uchimoto[†] **Yujie Zhang**[†] **Kiyoshi Sudo**[‡]

Masaki Murata[†] **Satoshi Sekine**[‡] **Hitoshi Isahara**[†]

[†]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun,
Kyoto 619-0289, Japan
{uchimoto,yujie,murata,isahara}@nict.go.jp

[‡]New York University
715 Broadway, 7th floor
New York, NY 10003, USA
{sudo,sekine}@cs.nyu.edu

Abstract

This paper describes Japanese-English-Chinese aligned parallel treebank corpora of newspaper articles. They have been constructed by translating each sentence in the Penn Treebank and the Kyoto University text corpus into a corresponding natural sentence in a target language. Each sentence is translated so as to reflect its contextual information and is annotated with morphological and syntactic structures and phrasal alignment. This paper also describes the possible applications of the parallel corpus and proposes a new framework to aid in translation. In this framework, parallel translations whose source language sentence is similar to a given sentence can be semi-automatically generated. In this paper we show that the framework can be achieved by using our aligned parallel treebank corpus.

1 Introduction

Recently, accurate machine translation systems can be constructed by using parallel corpora (Och and Ney, 2000; Germann et al., 2001). However, almost all existing machine translation systems do not consider the problem of translating a given sentence into a natural sentence reflecting its contextual information in the target language. One of the main reasons for this is that we had many problems that had to be solved by one-sentence to one-sentence machine translation before we could solve the contextual problem. Another reason is that it was difficult to simply investigate the influence of the context on the translation because sentence correspondences of the existing bilingual documents are rarely one-to-one, and are usually one-to-many or many-to-many.

On the other hand, high-quality treebanks such as the Penn Treebank (Marcus et al., 1993) and the Kyoto University text corpus (Kurohashi and Nagao, 1997) have contributed to improving the accuracies of fundamental techniques for natural language processing such as morphological analysis and syntactic structure analysis. However, almost all of these high-quality treebanks are based on monolingual corpora and do not have bilingual or multilingual information. There are few high-quality bilingual or multilingual treebank corpora because parallel corpora have mainly been actively used for machine translation between related languages such as English and French, therefore their syntactic structures are not required so much for aligning words or phrases. However, syntactic structures are necessary for machine translation between languages whose syntactic structures are different from each other, such as in Japanese-English, Japanese-Chinese, and Chinese-English machine translations, because it is more difficult to automatically align words or phrases between two unrelated languages than between two related languages. Actually, it has been reported that syntactic structures contribute to improving the accuracy of word alignment between Japanese and English (Yamada and Knight, 2001). Therefore, if we had a high-quality parallel treebank corpus, the accuracies of machine translation between languages whose syntactic structures are different from each other would improve. Furthermore, if the parallel treebank corpus had word or phrase alignment, the accuracy of automatic word or phrase alignment would increase by using the parallel treebank corpus as training data. However, so far, there is no aligned parallel treebank corpus whose domain is not restricted. For example, the Japanese Electronics Industry Development Association's (JEIDA's) bilingual corpus (Isahara and Haruno, 2000) has sentence, phrase, and proper noun alignment. However, it does not have morphological and syntactic information, the alignment is partial, and the target is restricted to a white paper. The Advance Telecommunications Research dialogue database (ATR, 1992) is a parallel treebank corpus between Japanese and English. However, it does not have word or phrase alignment, and the target domain is restricted to travel conversation.

Therefore, we have been constructing aligned parallel treebank corpora of newspaper articles between languages whose syntactic structures are different from each other since 2001; they

meet the following conditions.

1. It is easy to investigate the influence of the context on the translation, which means the sentences that come before and after a particular sentence, and that help us to understand the meaning of a particular word such as a pronoun.
2. The annotated information in the existing monolingual high-quality treebanks can be utilized.
3. They are open to the public.

To construct parallel corpora that satisfy these conditions, each sentence in the Penn Treebank (Release 2) and the Kyoto University text corpus (Version 3.0) has been translated into a corresponding natural sentence reflecting its contextual information in a target language by skilled translators, revised by native speakers, and each parallel translation has been annotated with morphological and syntactic structures, and phrasal alignment. Henceforth, we call the parallel corpus that is constructed by pursuing the above policy an *aligned parallel treebank corpus reflecting contextual information*. In this paper, we describe an aligned parallel treebank corpus of newspaper articles between Japanese, English, and Chinese, and its applications.

2 Construction of Aligned Parallel Treebank Corpus Reflecting Contextual Information

2.1 Human Translation of Existing Monolingual Treebank

The Penn Treebank is a tagged corpus of Wall Street Journal material, and it is divided into 24 sections. The Kyoto University text corpus is a tagged corpus of the *Mainichi* newspaper, which is divided into 16 sections according to the categories of articles such as the sports section and the economy section. To maintain the consistency of expressions in translation, a few particular translators were assigned to translate articles in a particular section, and the same translator was assigned to the same section. The instructions to translators for Japanese-English translation is basically as follows.

1. One-sentence to one-sentence translation as a rule

   Translate a source sentence into a target sentence. In case the translated sentence becomes unnatural by pursuing this policy, leave a comment.
2. Natural translation reflecting contextual information

   Except in the case that the translated sentence becomes unnatural by pursuing policy 1, translate a source sentence into a target sentence naturally.

   By deletion, replacement, or supplementation, let the translated sentence be natural in the context.

   In an entire article, the translated sentences must maintain the same meaning and information as those of the original sentences.
3. Translations of proper nouns

   Find out the translations of proper nouns by looking up the nouns in a dictionary or by using a web search. In case a translation cannot be found, use a temporary name and report it.

We started the construction of Japanese-Chinese parallel corpus in 2002. The Japanese sentences of the Kyoto University text corpus were also translated into Chinese by human translators. Then each translated Chinese sentence was revised by a second Chinese native. The instruction to the translators is the same as that given in the Japanese-English human translations.

The breakdown of the parallel corpora is shown in Table 1. We are planning to translate the remaining 18,714 sentences of the Kyoto University text corpus and the remaining 30,890 sentences of the Penn Treebank. As for the naturalness of the translated sentences, there are 207 (1%) unnatural English sentences of the Kyoto University text corpus, and 462 (2.5%) unnatural Japanese sentences of the Penn Treebank generated by pursuing policy 1.

2.2 Morphological and Syntactic Annotation

In the following sections, we describe the annotated information of the parallel treebank corpus based on the Kyoto University text corpus.

2.2.1 Morphological and Syntactic Information of Japanese-English corpus

Translated English sentences were analyzed by using the Charniak Parser (Charniak, 1999). Then, the parsed sentences were manually revised. The definitions of part-of-speech (POS) categories and syntactic labels follow those of the Treebank I style (Marcus et al., 1993). We have finished revising the 10,328 parsed sentences that appeared from January 1st to 11th. An example of morphological and syntactic structures is shown in Figure 1. In this figure, "S-ID" means the sentence ID in the Kyoto University text corpus. EOJ means the boundary between a Japanese parsed sentence and an English parsed sentence. The definition of Japanese morphological and syntactic information follows that of the Kyoto University text corpus (Version 3.0). The syntactic structure is represented by dependencies between Japanese phrasal units called *bunsetsu*s. The phrasal

Table 1: Breakdown of the parallel corpora

Original corpus	Languages	# of parallel sentences
Kyoto University text corpus	Japanese-English	19,669 (from Jan. 1st to 17th in 1995)
	Japanese-Chinese	38,383 (all)
Penn Treebank	Japanese-English	18,318 (from section 0 to 9)
Total	Japanese-English	37,987 (Approximately 900,000 English words)
	Japanese-Chinese	38,383 (Approximately 900,000 Chinese words)

```
# S-ID:950104141-008
* 0 2D
いずれも いずれも * 副詞 * * *
* 1 2D
十九 じゅうきゅう * 名詞 数詞 * *
歳 さい * 接尾辞 名詞性名詞助数辞 * *
前後 ぜんご * 接尾辞 名詞性名詞接尾辞 * *
の の * 助詞 接続助詞 * *
* 2 6D
若者 わかもの * 名詞 普通名詞 * *
で で だ 判定詞 * 判定詞 ダ列タ系連用テ形
、 、 * 特殊 読点 * *
* 3 4D
質問 しつもん * 名詞 サ変名詞 * *
に に * 助詞 格助詞 * *
* 4 5D
答える こたえる 答える 動詞 * 母音動詞 基本形
* 5 6D
気力 きりょく * 名詞 普通名詞 * *
も も * 助詞 副助詞 * *
* 6 -1D
残って のこって 残る 動詞 * 子音動詞ラ行 タ系連用テ形
い い いる 接尾辞 動詞性接尾辞 母音動詞 未然形
ない ない ない 接尾辞 形容詞性述語接尾辞 イ形容詞アウオ段 基本形
。 。 * 特殊 句点 * *
EOJ
(S1 (S (NP (PRP They))
       (VP (VP (VBD were)
               (NP (DT all))
               (ADJP (NP (QP (RB about)
                             (CD nineteen))
                         (NNS years))
                     (JJ old)))
           (CC and)
           (VP (VBD had)
               (S (NP (DT no)
                      (NN strength))
                  (VP (VBN left)
                      (SBAR (S (VP (ADVP (RB even))
                                   (TO to)
                                   (VP (VB answer)
                                       (NP (NNS questions))))))))))
       (. .)))
EOE
```

Figure 1: Example of morphological and syntactic information.

units or *bunsetsu*s are minimal linguistic units obtained by segmenting a sentence naturally in terms of semantics and phonetics, and each of them consists of one or more morphemes.

2.2.2 Chinese Morphological Information of Japanese-Chinese corpus

Chinese sentences are composed of strings of Hanzi and there are no spaces between words. The morphological annotation, therefore, includes providing tags of word boundaries and POSs of words. We analyzed the Chinese sentences by using the morphological analyzer developed by Peking University (Zhou and Duan, 1994). There are 39 categories in this POS set. Then the automatically tagged sentences were revised by the third native Chinese. In this pass the Chinese translations were revised again while the results of word segmentation and POS tagging were revised. Therefore the Chinese translations are obtained with a high quality. We have finished revising the 12,000 tagged sentences. The revision of the remaining sentences is ongoing. An example of tagged Chinese sentences is shown in Figure 2. The letters shown

```
S-ID:950104141-008
这些(ZheXie)/r
俄军(EJun)/j
士兵(ShiBing)/n
均(Jun)/d
为(Wei)/v
十九(ShiJiu)/m
岁(Sui)/q
左右(ZuoYou)/m
的(De)/u
年青人(NianQingRen)/n
,/w
他们(TaMen)/r
甚至(ShenZhi)/d
连(Lian)/p
回答(HuiDa)/v
问题(WenTi)/n
的(De)/u
气力(QiLi)/n
也(Ye)/d
没有(MeiYou)/v
。/w
```

Figure 2: Example of morphological information of Chinese corpus.

after '/' indicate POSs. The Chinese sentence is the translation of the Japanese sentence in Figure 1. The Chinese sentences are GB encoded. The 38,383 translated Chinese sentences have 1,410,892 Hanzi and 926,838 words.

2.3 Phrasal Alignment

This section describes the annotated information of 19,669 sentences of the Kyoto University text corpus.

The minimum alignment unit should be as small as possible, because bigger units can be constructed from units of the minimum size. However, we decided to define a *bunsetsu* as the minimum alignment unit. One of the main reasons for this is that the smaller the unit is, the higher the human annotation cost is. Another reason is that if we define a word or a morpheme as a minimum alignment unit, expressions such as post-positional particles in Japanese and articles in English often do not have alignments. To

effectively absorb those expressions and to align as many parts as possible, we found that a bigger unit than a word or a morpheme is suitable as the minimum alignment unit. We call the minimum alignment based on *bunsetsu* alignment units the *bunsetsu unit translation pair*. Bigger pairs than the *bunsetsu* unit translation pairs can be automatically extracted based on the *bunsetsu* unit translation pairs. We call all of the pairs, including *bunsetsu* unit translation pairs, *translation pairs*. The *bunsetsu* unit translation pairs for idiomatic expressions often become unnatural. In this case, two or more *bunsetsu* units are combined and handled as a minimum alignment unit. The breakdown of the *bunsetsu* unit translation pairs is shown in Table 2.

Table 2: Breakdown of the *bunsetsu* unit translation pairs.

(1) total # of translation pairs	172,255
(2) # of different translation pairs	146,397
(3) # of Japanese expressions	110,284
(4) # of English expressions	111,111
(5) average # of English expressions corresponding to a Japanese expression	1.33 ((2)/(3))
(6) average # of Japanese expressions corresponding to a English expression	1.32 ((2)/(4))
(7) # of ambiguous Japanese expressions	15,699
(8) # of ambiguous English expressions	12,442
(9) # of *bunsetsu* unit translation pairs consisting of two or more *bunsetsu*s	17,719

An example of phrasal alignment is shown in Figure 3. A Japanese sentence is shown from the line after the S-ID to the EOJ. Each line indicates a *bunsetsu*. Each rectangular line indicates a dependency between *bunsetsu*s. The leftmost number in each line indicates the *bunsetsu* ID. The corresponding English sentence is shown in the next line after that of the EOJ (End of Japanese) until the EOE (End of English). The English expressions corresponding to each *bunsetsu* are tagged with the corresponding *bunsetsu* ID such as <P id="*bunsetsu* ID"></P>. When there are two or more figures in the tag id such as id="1,2", it means two or more *bunsetsu*s are combined and handled as a minimum alignment unit.

For example, we can extract the following translation pairs from Figure 3.

- (J) 輸入が (*yunyuu-ga*) / 解禁された (*kaikin-sa-reta*); (E)that had been under the ban
- (J) 米国産リンゴの (*beikoku-san-ringo-no*); (E)of apples imported from the U.S.
- (J) 第１便が (*dai-ichi-bin-ga*); (E)The first cargo
- (J) 売り出された。(*uridasa-reta*); (E)was brought to the market.
- (J) 米国産リンゴの (*beikoku-san-ringo-no*) / 第１便が (*dai-ichi-bin-ga*); (E)The first cargo / of apples imported from the U.S.

```
# S-ID:950110003-001
1       輸入が┐
2     解禁された┐
3   米国産リンゴの┐
4           第１便が──┐
5              ９日、──┤
6           検疫手続きを┤
7                 終え、────P
8                 首都圏の┐   │
9           大手スーパーなどで──┤
10                      初めて┤
11                  売り出された。
EOJ
<P id="4">The first cargo</P> <P id="3">of apples
imported from the U.S.</P> <P id="1,2">that had been
under the ban</P> <P id="7">completed</P> <P id="6">
quarantine</P> <P id="7">and</P> <P id="11">was brought
to the market</P> <P id="10">for the first time</P>
<P id="5">on the 9th</P> <P id="9">at major supermarket
chain stores</P> <P id="8">in the Tokyo metropolitan
area</P> <P id="11">.</P>
EOE
```

Figure 3: Example of phrasal alignment.

- (J) 米国産リンゴの (*beikoku-san-ringo-no*) / 第１便が (*dai-ichi-bin-ga*) / 売り出された。(*uridasa-reta*); (E)The first cargo / of apples imported from the U.S. / was brought to the market.

Here, Japanese and English expressions are divided by the symbol ";", and "/" means a *bunsetsu boundary*.

An overview of the criteria of the alignment is as follows. Align as many parts as possible, except if a certain part is redundant. More detailed criteria will be attached with our corpus when it is open to the public.

1. Alignment of English grammatical elements that are not expressed in Japanese

   English articles, possessive pronouns, infinitive *to*, and auxiliary verbs are joined with nouns and verbs.
2. Alignment between a noun and its substitute expression

   A noun can be aligned with its substitute expression such as a pronoun.
3. Alignment of Japanese ellipses

   An English expression is joined with its related elements. For example, the English subject is joined with its related verb.
4. Alignment of supplementary or explanatory expression in English

   Supplementary or explanatory expressions in English are joined with their related words.

■ Ex. :

```
# S-ID:950104142-003
1    「佳」には──────┐
2      「美しい」P     │
3      「立派な」と┐   │
4              いう┐ │
5              意味が┤
6                ある。
EOJ
<P id="1">The Chinese character used for "ka"</P>
```

```
has such meanings as "beautiful" and "splendid."
EOE
・"「佳 (ka)」には (niwa)" corresponds to
  "The Chinese character used for "ka""
```

5. Alignment of date and time

   When a Japanese noun representing date and time is adverbial, the English preposition is joined with the date and time.
6. Alignment of coordinate structures

   When English expressions represented by "X (A + B)" correspond to Japanese expressions represented by "XA + XB", the alignment of X overlaps.

```
■ Ex. :
# S-ID:950106149-005
1   近畿圏では―――――――――┐
2       尼崎沖で I          |
3   八九年度から、―P        |
4         泉大津沖で I      |
5       九一年度から、――――┤
6           廃棄物などの┐   |
7                 投棄が┤
8               始まった。
EOJ
In the Kinki Region, disposal of wastes started
<P id="2"><P id="4"> at offshore sites of</P>
Amagasaki</P> and <P id="4">Izumiotsu</P> from
1989 and 1991 respectively.
EOE
・"尼崎沖 (Amagasaki-oki) で (de)" corresponds to
"at offshore sites of Amagasaki"
・"泉大津沖 (Izumiotsu-oki) で (de)" corresponds to
"at offshore sites of … Izumiotsu"
```

3 Applications of Aligned Parallel Treebank Corpus

3.1 Use for Evaluation of Conventional Methods

The corpus as described in Section 2 can be used for the evaluation of English-Japanese and Japanese-English machine translation. We can directly compare various methods of machine translation by using this corpus. It can be summarized as follows in terms of the characteristics of the corpus.

One-sentence to one-sentence translation can be simply used for the evaluation of various methods of machine translation.

Morphological and syntactic information can be used for the evaluation of methods that actively use morphological and syntactic information, such as methods for example-based machine translation (Nagao, 1981; Watanabe et al., 2003), or transfer-based machine translation (Imamura, 2002).

Phrasal alignment is used for the evaluation of automatically acquired translation knowledge (Yamamoto and Matsumoto, 2003).

An actual comparison and evaluation is our future work.

3.2 Analysis of Translation

One-sentence to one-sentence translation reflects contextual information. Therefore, it is suitable to investigate the influence of the context on the translation. For example, we can investigate the difference in the use of demonstratives and pronouns between English and Japanese. We can also investigate the difference in the use of anaphora.

Morphological and syntactic information and phrasal alignment can be used to investigate the appropriate unit and size of translation rules and the relationship between syntactic structures and phrasal alignment.

3.3 Use in Conventional Systems

One-sentence to one-sentence translation can be used for training a statistical translation model such as GIZA++ (Och and Ney, 2000), which could be a strong baseline system for machine translation.

Morphological and syntactic information and phrasal alignment can be used to acquire translation knowledge for example-based machine translation and transfer-based machine translation.

In order to show what kind of units are helpful for example-based machine translation, we investigated whether the Japanese sentences of newspaper articles appearing on January 17, 1995, which we call test-set sentences, could be translated into English sentences by using translation pairs appearing from January 1st to 16th as a database. First, we found that only one out of 1,234 test-set sentences agreed with one out of 18,435 sentences in the database. Therefore, a simple sentence search will not work well. On the other hand, 6,659 *bunsetsu*s out of 12,632 *bunsetsu*s in the test-set sentences agreed with those in the database. If words in *bunsetsu*s are expanded into their synonyms, the combination of the expanded *bunsetsu*s sets in the database may cover the test-set sentences. Next, therefore, we investigated whether the Japanese test-set sentences could be translated into English sentences by simply combining translation pairs appearing in the database. Given a Japanese sentence, words were extracted from it and translation pairs that include those words or their synonyms, which were manually evaluated, were extracted from the database. Then, the English sentence was manually generated by just combining English expressions in the extracted translation pairs. One hundred two relatively short sentences (the average number of *bunsetsu*s is about 9.8) were selected as inputs. The number of equivalent translations, which mean that the translated sentence is grammatical and has the same meaning as the source

sentence, was 9. The number of similar translations, which mean that the translated sentence is ungrammatical, or different or wrong meanings of words, tenses, and prepositions are used in the translated sentence, was 83. The number of other translations, which mean that some words are missing, or the meaning of the translated sentence is completely different from that of the original sentence, was 10. For example, the original parallel translation is as follows:

Japanese: さきがけ側は通常国会に向け、政策や国会運営をテーマとする協議機関を両党に設置することを提案した。
English: New Party Sakigake proposed that towards the ordinary session, both parties found a council to discuss policy and Diet management.

Given the Japanese sentence, the translated sentence was:

Translation: Sakigake Party suggested to set up an organization between the two parties towards the regular session of the Diet to discuss under the theme of policies and the management of the Diet.

This result shows that only 9% of input sentences can be translated into sentences equivalent to the original ones. However, we found that approximately 90% of input sentences can be translated into English sentences that are equivalent or similar to the original ones.

3.4 Similar Parallel Translation Generation

The original aim of constructing an aligned parallel treebank corpus as described in Section 2 is to achieve a new framework for translation aid as described below.

It would be very convenient if multilingual sentences could be generated by just writing sentences in our mother language. Today, it can be formally achieved by using commercial machine translation systems. However, the automatically translated sentences are often incomprehensible. Therefore, we have to revise the original and translated sentences by finding and referring to parallel translation whose source language sentence is similar to the original one. In many cases, however, we cannot find such similar parallel translations to the input sentence. Therefore, it is difficult for users who do not have enough knowledge of the target languages to generate comprehensible sentences in several languages by just searching similar parallel translations in this way. Therefore, we propose to generate similar parallel translations whose source language sentence is similar to the input sentence. We call this framework for translation aid *similar parallel translation generation*.

We investigated whether the framework can be achieved by using our aligned parallel treebank corpus. As the first step of this study, we investigated whether an appropriate parallel translation can be generated by simply combining translation pairs extracted from our aligned parallel treebank corpus in the following steps.

1. Extract each content word with its adjacent function word in each *bunsetsu* in a given sentence
2. The extracted content words and their adjacent function words are expanded into their synonyms and class words whose major and minor POS categories are the same
3. Find translation pairs including the expanded content words with their expanded adjacent function words in the given sentence
4. For each *bunsetsu*, select a translation pair that has similar dependency relationship to those in the given sentence
5. Generate a parallel translation by combining the selected translation pairs

The input sentences were randomly selected from 102 sentences described in Section 3.3. The above steps, except the third step, were basically conducted manually. The Examples of the input sentences and generated parallel translations are shown in Figure 4.

The basic unit of translation pairs in our aligned parallel treebank corpus is a *bunsetsu*, and the basic unit in the selection of translation pairs is also a *bunsetsu*. One of the advantages of using a *bunsetsu* as a basic unit is that a Japanese expression represented as one of various expressions in English, or omitted in English, such as Japanese post-positional particles, is paired with a content word. Therefore, the translation of such an expression is appropriately selected together with the translation of a content word when a certain translation pair is selected. If the translation of such an expression was selected independently of the translation of a content word, the combination of each translation would be ungrammatical or unnatural. Another advantage of the basic unit, *bunsetsu*, is that we can easily refer to dependency information between *bunsetsu*s when we select an appropriate translation pair because the original treebank has the dependency information between *bunsetsu*s. These advantages are utilized in the above generation steps. For example, in the first step, a content word "国会 (*kokkai*, Diet session)" in the second example in Figure 4 was extracted from the *bunsetsu* "通常国会 (*tsuujo-kokkai*, the ordinary Diet session) に (*ni*, case marker)", and it was expanded into its class word "会 (*kai*, meeting)" in the second step. Then, a translation pair "(J) 国連子どもの権利委員会 (*kokuren-kodomo-no-kenri-iinkai*) に (*ni*, case marker); (E)the UN Committee on the Rights of the Child /(J) 対し (*taishi*); (E)towards" was extracted as a translation pair in the third step. Since the dependency between "国連子どもの権利委員会

(*kokuren-kodomo-no-kenri-iinkai*, the UN Committee on the Rights of the Child)" and "対し (*taishi*, towards)" is similar to that between "通常国会 (*tsuujo-kokkai*, the ordinary Diet session) に (*ni*, case marker)" and "向け (*muke*, towards)" in the input sentence, this translation pair was selected in the fourth step. Finally, the *bunsetsu* "国連子どもの権利委員会 (*kokuren-kodomo-no-kenri-iinkai*, the UN Committee on the Rights of the Child) に (*ni*, case marker)" and its translation "the UN Committee on the Rights of the Child" was used for generation of a parallel translation in the fifth step.

When we use the generated parallel translation for the exact translation of the input sentence, we should replace "国連子どもの権利委員会 (*kokuren-kodomo-no-kenri-iinkai*)" and its translation "the UN Committee on the Rights of the Child" with "通常国会 (*tsuujo-kokkai*, the ordinary Diet session)" and its translation "the ordinary Diet session" by consulting a bilingual dictionary. In this example, "その (*sono*)" and "them" should also be replaced with "両党 (*ryoto*)" and "both parties". It is easy to identify words in the generated translation that should be replaced with words in the input sentence because each *bunsetsu* in translation pairs is already aligned. In such cases, templates such as "[会議 (*kaigi*)] に (*ni*) 向け (*muke*)" and "towards [council]" can be automatically generated by generalizing content words expanded in the second step and their translation in the generated translation. The average number of English expressions corresponding to a Japanese expression is 1.3 as shown in Table 2. Even when there are two or more possible English expressions, an appropriate English expression can be chosen by selecting a Japanese expression by referring to dependencies in extracted translation pairs. Therefore, in many cases, English sentences can be generated just by reordering the selected expressions. The English word order was estimated manually in this experiment. However, we can automatically estimate English word order by using a language model or an English surface sentence generator such as FERGUS (Bangalore and Rambow, 2000). Unnatural or ungrammatical parallel translations are sometimes generated in the above steps. However, comprehensible translations can be generated as shown in Figure 4. The biggest advantage of this framework is that comprehensible target sentences can be generated basically by referring only to source sentences. Although it is costly to search and select appropriate translation pairs, we believe that human labor can be reduced by developing a human interface. For example, when we use a Japanese text generation system from keywords (Uchimoto et al., 2002), users should only select appropriate keywords.

We are investigating whether or not we can generate similar parallel translations to all of the Japanese sentences appearing on January 17, 1995. So far, we found that we can generate similar parallel translations to 691 out of 840 sentences (the average number of *bunsetsu*s is about 10.3) including the 102 sentences described in Section 3.3. We found that we could not generate similar parallel translations to 149 out of 840 sentences.

In the proposed framework of similar parallel translation generation, the language appearing in a corpus corresponds to a controlled language, and users are allowed to use only the controlled language to write sentences in the source language. We believe that high-quality bilingual or multilingual documents can be generated by letting us adapt ourselves to the controlled environment in this way.

4 Conclusion

This paper described aligned parallel treebank corpora of newspaper articles between languages whose syntactic structures are different from each other; they meet the following conditions.

1. It is easy to investigate the influence of the context on the translation.
2. The annotated information in the existing monolingual high-quality treebanks can be utilized.
3. It is open to the public.

To construct parallel corpora that satisfy these conditions, each sentence in the existing monolingual high-quality treebanks has been translated into a corresponding natural sentence reflecting its contextual information in a target language by skilled translators, and each parallel translation has been annotated with morphological and syntactic structures and phrasal alignment.

This paper also described the possible applications of the parallel corpus and proposed a similar parallel translation generation framework. In this framework, a parallel translation whose source language sentence is similar to a given sentence can be semi-automatically generated. In this paper we demonstrated that the framework could be achieved by using our aligned parallel treebank corpus.

In the near future, the aligned parallel treebank corpora will be open to the public, and expanded. We are planning to use the corpora actively for machine translation, as a translation aid, and for second language learning. We are also planning to develop automatic or semi-automatic alignment system and an efficient interface for machine translation aid.

Input sentence (Japanese only)	村山首相と武村蔵相との会談は首相公邸で行われ、新民主連合の新会派結成を中心に意見交換。(Prime Minister Murayama and Finance Minister Takemura met in the presidential office and they exchanged their opinions, mainly on the issue of the new faction being formed by the New Democratic Union.)
Generated parallel translation	(J) 村山富市首相と武村蔵相は会談を首相公邸で開き、新民主連合の新党結成を中心に意見交換した。 (E) Finance Minister Takemura held the meeting at the official residence to exchange views about the formation of the new party of the New Democratic Union.
Input sentence (Japanese only)	さきがけ側は通常国会に向け、政策や国会運営をテーマとする協議機関を両党に設置することを提案した。(New Party Sakigake proposed that towards the ordinary session, both parties found a council to discuss policy and Diet management.)
Generated parallel translation	(J) さきがけは国連子どもの権利委員会に向け、政策や国会運営をテーマに、話し合う機関をその間に設置することを提案した (E) Sakigake proposed to set up an organization between them towards the UN Committee on the Rights of the Child to discuss under the theme of policies and the management of the Diet.
Input sentence (Japanese only)	会談は新民連による新党が新進党との連携を深める形で進んでいることをけん制する意味もあった。(The meeting was also intended to slow the movement towards the new party by the New Democratic Union, which is trying to deepen the relationship with the New Frontier Party.)
Generated parallel translation	(J) 会談は新民連による新党が新進党との連携を深める形に進んでいることをけん制した意味があった。 (E) The meeting had meanings to restrict the movement that the new party of New Democratic Union is progressing to strengthen the coalition with The New Frontier Party.
Input sentence (Japanese only)	新進党の川端達夫衆院議員は十六日、山花貞夫氏らとの新会派結成のため、十七日に同党に離党届を提出することを決めた。(Lower House Diet Member Tatsuo Kawabata of the New Frontier Party decided on the 16th that he would hand in notification of his secession to the party on the 17th, in order to form a new faction with Sadao Yamahana's group.)
Generated parallel translation	(J) 新進党の川端達夫衆院議員は、十六日、天野祐吉氏らと新会派結成のため、十七日に新生党に離党届は提出することを決めた。 (E) On 16th Tatsuo Kawabata, a member of the House of Representatives of the New Frontier Party decided to submit The notice to leave the party to the Shinsei Party on the 17th in order to establish a new faction with Yuukichi Amano and others.
Input sentence (Japanese only)	参院会派名は民主改革連合との関係を詰めてから決定する。(As for the faction name in the Upper House, they will decide after they consider how to form a relationship with Democratic Reform Union.)
Generated parallel translation	(J) 会派名は連合との関係を話し合って決定する。 (E) The name of the faction will be decided after discussing the relationship with the JTUC.

Figure 4: Example of generated similar parallel translations.

Acknowledgments

We thank the Mainichi Newspapers for permission to use their data.

References

ATR. 1992. Dialogue Database. http://www.red.atr.co.jp/database_page/taiwa.html.

S. Bangalore and O. Rambow. 2000. Exploiting a Probabilistic Hierarchical Model for Generation. In *Proceedings of the COLING*, pages 42–48.

E. Charniak. 1999. A Maximum-Entropy-Inspired Parser. Technical Report CS-99-12.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada 2001. Fast Decoding and Optimal Decoding for Machine Translation. In *Proceedings of the ACL-EACL*, pages 228–235.

K. Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proceedings of the TMI*, pages 74–84.

H. Isahara and M. Haruno. 2000. Japanese-English aligned bilingual corpora. In Jean Veronis, editor, *Parallel Text Processing - Alignment and Use of Translation Corpora*, pages 313–334. Kluwer Academic Publishers.

S. Kurohashi and M. Nagao. 1997. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the NLPRS*, pages 451–456.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M. Nagao. 1981. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*.

F. J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the ACL*, pages 440–447.

K. Uchimoto, S. Sekine, and H. Isahara. 2002. Text Generation from Keywords. In *Proceedings of the COLING*, pages 1037–1043.

H. Watanabe, S. Kurohashi, and E. Aramaki. 2003. Finding Translation Patterns from Paired Source and Target Dependency Structures. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 397–420. Kluwer Academic Publishers.

K. Yamada and K. Knight. 2001. A Syntax-based Statistical Translation Model. In *Proceedings of the ACL*, pages 523–530.

K. Yamamoto and Y. Matsumoto. 2003. Extracting Translation Knowledge from Parallel Corpora. In Michael Carl and Andy Way, editors, *Recent Advances in Example-Based Machine Translation*, pages 365–395. Kluwer Academic Publishers.

Q. Zhou and H. Duan. 1994. Segmentation and POS Tagging in the Construction of Contemporary Chinese Corpus. *Journal of Computer Science of China*, Vol.85. (in Chinese)

JMdict: a Japanese-Multilingual Dictionary

James BREEN
Monash University
Clayton 3800, Australia
jwb@csse.monash.edu.au

Abstract

The JMdict project has at its aim the compilation of a multilingual lexical database with Japanese as the pivot language. Using an XML structure designed to cater for a mix of languages and a rich set of lexicographic information, it has reached a size of approximately 100,000 entries, with most entries having translations in English, French and German. The compilation involves information re-use, with the French and German translations being drawn from separately maintained lexicons. Material from other languages is also being included. The file is freely available for research purposes and for incorporation in dictionary application software, and is available in several WWW server systems.

1 Introduction

The JMdict project has as its primary goal the compilation of a Japanese-multilingual dictionary, i.e. a dictionary in which the headwords are from the Japanese lexicon, and the translations are in several other languages. It may be viewed as a synthesis of a series of Japanese-Other Language bilingual dictionaries, although, as discussed below, there is merit in having this information collocated.

The project grew out of, and has now subsumed, an earlier Japanese-English dictionary project (EDICT: Electronic Dictionary) (Breen, 1995, 2004a). With Japanese being an important language in world trade, and with it being the second most common language used on the WWW, it is not surprising that there is considerable interest in electronic lexical resources for Japanese in combination with other languages.

2 Project Goals and Development

As mentioned above, the JMdict project grew out of the bilingual EDICT dictionary project. The EDICT project began in the early 1990s with a relatively simple goal of producing a Japanese-English dictionary file that could be used in basic software packages to provide traditional dictionary services, as well as facilities to assist reading Japanese text. The format was (and is) quite simple, comprising lines of text consisting of a Japanese word written using kanji and/or kana, the reading (pronunciation) of that word in kana, and one or more English translations.

By the late 1990s, the file had outgrown its humble origins, having reached over 50,000 entries, and having spun off a parallel project for recording Japanese proper nouns (see below). The material has partly been drawn from word lists, vocabulary lists, etc. in the public domain, and supplemented by material prepared by large numbers of users and other volunteers wishing to contribute. While it had been used in a variety of software systems, and as a source of lexical material in a number of projects, it was clear that its structure was quite inadequate for the lexical demands being made by users. In particular, it was not able to incorporate a suitable variety of information, nor represent the orthographical complexities of the source language. Accordingly, in 1999 it was decided to launch a new dictionary

project incorporating the information from the EDICT file, but expanded to include translations from other languages with the Japanese entries remaining as the pivots. The project goals were:

a. a file format, preferably using a recognized standard, which would enable ready access and parsing by a variety of software applications;

b. the handling of orthographical and pronunciation variation within the single entry. This addressed a major problem with the EDICT format, as many Japanese words can be written with alternative kanji and with varying portions in kana *(okurigana),* and may have alternative pronunciations. The EDICT format required each variant to be treated as a separate entry, which added to the complexity of maintaining and extending the dictionary;

c. additional and more appropriately associated tagging of grammatical and other information. Certain information such as the part of speech or the source language of loan words had been added to the EDICT file in parentheses within the translation fields, but the scope was limited and the information could not easily be parsed;

d. provision for differentiation between different senses in the translations. While basic indication of polysemy had been provided in the EDICT file by prepending (1), (2), etc. to groups of translations, the result was difficult to parse. Also it did not support the case where a sense or nuance was tied to a particular pronunciation, as occurs occasionally in Japanese;

e. provision for the inclusion of translational equivalents from several languages. The EDICT dictionary file was being used in a number of countries, and several informal projects had begun to develop equivalent files for Japanese and other target languages. A small Japanese-German file (JDDICT) had been released in the EDICT format. There was considerable interest expressed in having translations in various languages collocated to enable such things as having a single reference file for several languages, cross-referencing of entries, cross-language retrieval, etc. as well as acting as a focus for the possible development of translations for as yet unrepresented languages;

f. provision for inclusion of examples of the usage of words. As the file expanded, many users of the file requested some form of usage examples to be associated with the words in the file. The EDICT format was not capable of supporting this;

g. provision for cross-references to related entries;

h. continued generation of EDICT-format files. As a large number of packages and servers had been built around the EDICT format, continued provision of content in this format was considered important, even if the information only contained a sub-set of what was available.

An early decision was to use XML (Extensible Markup Language) as a format for the JMdict file, as this was expected to provide the appropriate flexibility in format, and was also expected to be supported by applications, parsing libraries, etc.

An examination was made of other available dictionary formats to ascertain if a suitable formatting model was available. It was known that commercial dictionary publishers has well-structured databases of lexical information, and some were moving to XML, but none of the details were available. A large number of bilingual dictionary files and word lists were in the public domain; however in general they only used very simple structures, and none could be found which covered all the content requirements of the project. The dictionary section of the TEI (Text Encoding Initiative), which at the time of writing has a well-developed document structure for bilingual dictionaries, was at that stage quite limited (Sperberg-McQueen et al, 1999). Accordingly, an XML DTD (Document Type Definition) was developed which was tailored to the requirements of the project.

The EDICT file was parsed and reformatted into the JMdict structure, and at the same time, many of the orthographical variants were identified and merged. The initial release of the DTD and XML-format file took place in May 1999. At that stage, it contained the English translations from the EDICT file and the German translations from the JDDICT file. As described below, it has been expanded considerably since then, both in terms of number of entries and also in multi-lingual coverage.

3 Project Status

The JMdict file was first released in 1999, and updated versions are released 3-4 times each year along with versions of the EDICT file, which is generated at the same time from the same data files. The file now has over 99,300 entries, i.e. the size of a medium-large printed dictionary, and the growth in numbers of entries is now relatively slow, with most updates dealing with corrections and expansion of existing entries.

The file is available under a liberal licence that allows its use for almost any purpose without fee. The only requirement is that its use be fully acknowledged and that any files developed from it continue under the same licence conditions.

4 Structure

The JMdict XML structure contains one element type: **<entry>,** which in turn contains sequence number, kanji word, kana word, information and translation elements. The sequence number is used for maintenance and identification.

The kanji word and kana word elements contain the two forms of the Japanese headwords; the former is used for representations containing at least one kanji character, while the latter is for representations in kana alone. The kana word is effectively the pronunciation, but is also an important key for indexing the dictionary file, as Japanese dictionaries are usually ordered by kana words. The minimum content of these fields is a single word in the kana word element. In addition, each entry may contain information about the words (unusual orthographical variant, archaic kanji, etc.) and frequency of use information. The latter needs to be associated with the actual words rather than the entry as a whole because some combinations of kanji and kana words are used more frequently than others. (For example, 合気道 and 合氣道 are orthographical variants of the one word *(aikidô)*, but the former is more common.)

The kana used in the elements follows modern Japanese orthography, i.e. *hiragana* is used for native Japanese words, and *katakana* for loan words, onomatopoeic words, etc.

In most cases an entry has just one kanji and one kana word (approx. 75%), or one kana word alone (15%). In about 10% of entries there are multiple words in one of the elements. In some cases a kana reading can only be associated with a subset of the kanji words in the entry. For example, *soyokaze* (そよかぜ: breeze) can be written either 微風 or そよ風 (the latter is more common as そよ is a non-standard reading of the 微 kanji). However 微風 can also be pronounced *bifuu* (びふう) with the same meaning, but clearly this pronunciation cannot be associated with the そよ風 form, as the kana portion is read *"soyo"*. XML does not provide an elegant method for indicating a restricted mapping between portions of two elements, so when such a restriction is required, additional tags are used with each kana word supplying the kanji word with which it may be validly associated.

The information element contains general information about the Japanese word or the entry as a whole. The contents allow for ISO-639 source language codes (for loan

words), dialect codes, etymology, bibliographic information and update details.

The translation area consists of one or more sense elements that contain at a minimum a single gloss. Associated with each sense is a set of elements containing part of speech, cross-reference, synonym/antonym, usage, etc. information. Also associated with the sense may be restriction codes tying the sense to a subset of the Japanese words. For example, 水気 can be pronounced *suiki* (すいき) and *mizuge* (みずけ); both meaning "moisture", but the former alone can also mean "dropsy".

The gloss element has an attribute stating the target language of the translation. In its absence it is assumed the gloss is in English. There is also an attribute stating the gender, if for example, the part-of-speech is a noun and the gloss is in a language with gendered nouns. Figure 1 shows a slightly simplified example of an entry. The **<ke_pri>** and **<re_pri>** elements indicate the word is a member of a particular set of common words.

```
<entry>
<ent_seq>1206730</ent_seq>
<k_ele>
<keb>学校</keb>
<ke_pri>ichi1</ke_pri>
</k_ele>
<r_ele>
<reb>がっこう</reb>
<re_pri>ichi1</re_pri>
</r_ele>
<sense>
<pos>&n;</pos>
<gloss>school</gloss>
<gloss g_lang="nl" g_gend="fg">school</gloss>
<gloss g_lang="fr" g_gend="fg">école</gloss>
<gloss g_lang="ru" g_gend="fg">школа</gloss>
<gloss g_lang="de" g_gend="fg">Schule</gloss>
<gloss g_lang="de"
g_gend="fg">Lehranstalt</gloss>
</sense>
</entry>
```

Fig. 1: Example JMdict entry

The potential to have multiple kanji and kana words within an entry brings attention to the issues of homonymy, homography and polysemy, and the policies for handling these, in particular the criteria for combining kanji and kana words into a single entry. As Japanese has a comparatively limited set of phonemes there are a large number of homophonous words. For example, over twenty different words have the kana representation こうじょう *(kôjô)*. If we regard homography as only applying to words written wholly or partly with kanji, there are relatively few cases of it, however they do exist, e.g. 川柳 when read せんりゅう *(senryû)* means a comic poem, but when read かわやなぎ *(kawayanagi)* means a variety of willow tree.

The combining rule that has been applied in the compilation of the JMdict file is as follows:

a. treat each basic entry as a triplet consisting of: kanji representation, matching kana representation, senses;

b. if for any basic entries two or more members of the triplet are the same, combine them into the one entry;

   i. if the entries differ in kanji or kana representation, include these as alternative forms;

   ii. if the entries differ in sense, treat as a case of polysemy;

c. in other cases leave the entries separate.

This rule has been applied successfully in a majority of cases. The main problems arise where the meanings are similar or related, as in the case of the entries: (放す, はなす, to separate; to set free; to turn loose) and (離す, はなす, to part; to divide; to separate), where the kana words are the same and the meanings overlap. Japanese dictionaries are divided on 放す and 離す; some keeping them as separate entries, and others having them as the one entry with two

main senses. (The two words derive from a common source.)

5 Parts of Speech and Related Issues

As languages differ in their parts of speech (POS), the recording of those details in bilingual dictionaries can be a problem (Al-Kasimi, 1977). Traditionally bilingual dictionaries involving Japanese avoid recording any POS information, leaving it to the user to deduce that information from the translation and examples (if any). In the early stages of the EDICT project, POS information was deliberately kept to a minimum, e.g. indicating where a verb was transitive or intransitive when this was not apparent from the translation, mainly to conserve storage space. As there are a number of advantages in having POS information marked in an electronic dictionary file, a POS element was included in the JMdict structure, and publicly available POS classifications were used to populate much of the file. About 30% of entries remain to be classified; mostly nouns or short noun phrases.

In the interests of saving space an early decision had been made to avoid listing derived forms of words. For example, the Japanese adjective 高い *(takai)* meaning "high, tall, expensive" has derived forms of 高さ *(takasa)* "height" and 高く *(takaku)* "highly". As this process is very regular, many Japanese dictionaries do not carry entries for the derived forms, and some bilingual dictionaries follow suit. Another such example is the common verb form, sometimes called a "verbal noun", which is created by adding the verb する *(suru)* "to do" to appropriate nouns. The verb "to study" is 勉強する *(benkyôsuru)* where 勉強 is a noun meaning "study" in this context. Again, Japanese dictionaries often do not include these forms as headwords, preferring to indicate in the body of an entry that the formation is possible.

The omission of such derived forms means that care needs to be taken when constructing the translations so that the user is readily able to identify the appropriate translation of one of the derived forms.

In a multilingual context, the omission of derived forms can have other problems. The recording of する verbs only in their noun base form has been reported to lead to some discomfort among German users, as German language orthographical convention capitalizes the first letters of nouns but not verbs (the WaDokuJT file has する verbs as separate entries for this reason).

6 Inclusion and Maintenance of Multiple Languages

As mentioned above, part of the interest in having entries with translations in a range of languages came from the compilation of a number of dictionary files based on or similar to the EDICT file. There are a number of issues associated with the inclusion of material from other dictionary files, in particular those relating to the compilation policies: coverage, handling of inflected forms, etc. (Breen, 2002) There is also the major issue of the editing and maintenance of the material, which has the potential to become more complex as each language is incorporated.

The approach taken with JMdict has been to:

a. maintain a core Japanese-English file with a well-documented structure and set of inclusion and editing policies;
b. encourage the development and maintenance of equivalent files in other languages paired with Japanese, which can draw on the JMdict/EDICT material as required;
c. periodically build the complete multi-lingual JMdict from the different components.

This approach has proved successful in that it has separated the compilation of the file

from the ongoing editing of the components, and has left the latter in the hands of those with the skills and motivation to perform the task.

At the time of writing, the JMdict file has over 99,300 entries (Japanese and English), of which 83,500 have German translations, 58,000 have French translations, 4,800 have Russian translations and 530 have Dutch translations. A set of approximately 4,500 Spanish translations is being prepared, with the prospects that some 20,000 will be available shortly.

The major sources of these additional translations are:

a. French translations from two projects:

i. approximately 17,500 entries have come from the Dictionnaire français-japonais Project (Desperrier, 2002), a project to translate the most common Japanese words from the EDICT File into French;

ii. a further 40,500 entries drawn from the 仏語補完計画 (French-Japanese Complementation Project) at http://francais.sourceforge.jp/ (This project is also based on the EDICT file.)

b. German translations from the WaDokuJT Project (Apel, 2002). This is a large file of over 300,000 entries; however, unlike JMdict it includes many phrases, proper nouns and inflected forms of verbs, etc. The overlap of coverage with JMdict is quite high, leading to the large number of entries that have been included in the JMdict file.

One of the issues that can lead to problems when incorporating translations from other project files is that of aligning the translations when an entry has several senses. In the case of the French translations, the project coordinator has marked the translations of polysemous entries with a sense code, thus enabling the translations to be inserted correctly when compiling the final file. For other languages, the translations are being appended to the set English translations. The appropriate handling of multiple senses is an item of future work.

7 Examples of Word Usage

When the project was begun and the DTD designed, it was intended that sets of bilingual examples of usage of the entry words would be included. For this reason an <example> element was associated with each sense to allow for such example phrases, sentences, etc, to be included.

In practice, a quite different approach has been taken. With the availability since 2001 of a large corpus of parallel Japanese/English sentences (Tanaka, 2001), it was decided to keep the corpus intact, and instead provide for the association of selected sentences from the corpus with dictionary entries via dictionary application software (Breen, 2003b). This strategy, which required the corpus to be parsed to extract a set of index words for each sentence, has proved effective at the application level. It also has the advantage of decoupling the maintenance of the dictionary file from that of the example corpus.

8 Related Projects

Apart from a few small word lists involving several European languages, the only other major current project attempting to compile a comprehensive multilingual database is the Papillon project (e.g. Boitet et al, 2002). See http://www.papillon-dictionary.org/ for a full list of publications. The Papillon design involves linkages based on word-senses as proposed in (Sérasset, 1994) with the finer lexical structure based on Meaning-Text

Theory (MTT) (Mel'cuk, 1984-1996). At the time of writing the Papillon database is still in the process of being populated with lexical information.

Closely related to the JMdict project is the Japanese-Multilingual Named Entity Dictionary (JMnedict) project. This is a database of some 400,000 Japanese place and person names, and non-Japanese names in their Japanese orthographical form, along with a romanized transcription of the Japanese (Breen, 2004b). Some geographical names have English descriptions: cape, island, etc. which are in the process of being extended to other languages. The JMnedict file is in an XML format with a similar structure to JMdict.

Another multilingual lexical database is KANJIDIC2 (Breen, 2004c), which contains a wide range of information about the 13,039 kanji in the JIS X 0208, JIS X 0212 and JIS X 0213 character standards. Among the information for each kanji are the set of readings in Japanese, Chinese and Korean, and the broad meanings of each kanji in English, German and Spanish. A set of Portuguese meanings is being prepared. The database is in an XML format.

9 Applications

While there are a number of experimental systems using the JMdict file, the only application system using the full multilingual file at present is the Papillon project server. Figure 2 shows the display from that server when looking up the word 川柳. The author's WWWJDIC server (Breen, 2003a) uses the Japanese-English components of the file. Figure 3 is an extract from the WWWJDIC display for the word 小人, which is an example of an entry with multiple kana words, and senses restricted by reading. (The (P) markers indicate the more common readings.)

Fig. 2: Papillon example for 川柳

小人 【こびと(P); しょうにん(P); しょうじん; こども(rare)】 (n) (1) child; small person; (2) (こびと only) dwarf; (3) (しょうじん only) narrow-minded person; mean person

Fig. 3: WWWJDIC example for 小人

The EDICT Japanese-English dictionary file, which is generated from the same database as the JMdict file, continues to be a major non-commercial Japanese-English lexical resource, and is used in a large number of applications and servers, as well as in a number of research projects.

10 Conclusion

The JMdict project has successfully developed a multilingual lexical database using Japanese as the pivot language. In doing so, it has reached a lexical coverage comparable to medium-large printed dictionaries, and its components are used in a wide range of applications and research projects. It has also demonstrated the potential for re-use of material from related and cooperating lexicon projects. The files of the JMdict project are readily available for use by researchers and developers, and have the potential to be a significant lexical resource in a multilingual context.

References

Al-Kasami, A.M. 1977 *Linguistics and Bilingual Dictionaries,* E.J. Brill, Leiden

Apel, U. 2002. *WaDokuJT - A Japanese-German Dictionary Database,* Papillon 2002 Seminar, NII, Tokyo

Boitet, C, Mangeot-Lerebours, M, Sérasset, G. 2002 *The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons,* Proc. of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop, Ed. Wilcock, Ide & Romary, Taipei, Taiwan.

Breen, J.W. 1995. *Building an Electronic Japanese-English Dictionary,* JSAA Conference, Brisbane.

Breen, J.W. 2002. *Practical Issues and Problems in Building a Multilingual Lexicon,* Papillon 2002 Seminar, NII, Tokyo.

Breen, J.W. 2003a. *A WWW Japanese Dictionary,* in "Language Teaching at the Crossroads", Monash Asia Institute, Monash Univ. Press.

Breen, J.W. 2003b. *Word Usage Examples in an Electronic Dictionary,* Papillon 2003 Seminar, Sapporo.

Breen, J.W. 2004a. *The EDICT Project,* http://www.csse.monash.edu.au/~jwb/edict.html

Breen, J.W. 2004b. *The ENAMDICT/JMnedict Project,* http://www.csse.monash.edu.au/~jwb/enamdict_doc.html

Breen, J.W. 2004c. *The KANJIDIC2 Project,* http://www.csse.monash.edu.au/~jwb/kanjidic2/

Desperrier, J-M. 2002. *Analysis of the results of a collaborative project for the creation of a Japanese-French dictionary,* Papillon 2002 Seminar, NII, Tokyo.

Mel'cuk, I, et al. 1984-1996. *DEC: dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques,* Vols I-IV, Montreal Univ. Press.

Sérasset, G. 1994. *SUBLIM: un Système Universel de Bases Lexicales Multilingues et NADIA: sa spécialisation aux bases lexicales interlingues par acceptions,* (Doctoral Thesis) Joseph Fourier University, Grenoble

Sperberg-McQueen, C.M.. and Burnard, L. (eds.) 1999. *Guidelines for Electronic Text Encoding and Interchange.* Oxford Univ. Press.

Tanaka, Y. 2001. *Compilation of a Multilingual Parallel Corpus* PACLING 2001, Japan.

A Generic Collaborative Platform for Multilingual Lexical Database Development

Gilles SÉRASSET
GETA-CLIPS, IMAG, Université Joseph Fourier
BP 53 – 38041 Grenoble cedex 9 – France
Gilles.Serasset@imag.fr

Abstract

The motivation of the Papillon project is to encourage the development of freely accessible Multilingual Lexical Resources by way of online collaborative work on the Internet. For this, we developed a generic community website originally dedicated to the diffusion and the development of a particular acception based multilingual lexical database.

The generic aspect of our platform allows its use for the development of other lexical databases. Adapting it to a new lexical database is a matter of description of its structures and interfaces by way of XML files. In this paper, we show how we already adapted it to other very different lexical databases. We also show what future developments should be done in order to gather several lexical databases developers in a common network.

1 Introduction

In order to cope with information available in many languages, modern information systems need large, high quality and multilingual lexical resources. Building such a resource is very expensive. To reduce these costs, we chose to use the "collaborative" development paradigm already used with LINUX and other open source developments.

In order to develop such a specific multilingual lexical database, we built a Web platform to gather an Internet community around lexical services (accessing many online dictionaries, contributing to a rich lexical database, validating contributions from others, sharing documents, ...). Initially built for the Papillon project, this platform is generic and allows for the collaborative development of other lexical resources (monolingual, bilingual or multilingual) provided that such resources are described to the platform.

After presenting the Papillon project and platform, we will show how we may give access to many existing dictionaries, using an unified interface. Then, we will present the edition service, and detail how it may be customised to handle other very different dictionaries.

2 The Papillon project

2.1 Motivations

Initially launched in 2000 by a French-Japanese consortium, the Papillon project[1] (Sérasset and Mangeot-Lerebours, 2001) rapidly extended its original goal — the development of a rich French Japanese lexical database — to its actual goal — the development of an Acception based Multilingual Lexical Database (currently tackling Chinese, English , French, German, Japanese, Lao, Malay, Thai and Vietnamese).

This evolution was motivated in order to:

- reuse many existing lexical resources even the ones that do not directly involve both initial languages,
- be reusable by many people on the Internet, hence raising the interest of others in its development,
- allow for external people (translator, native speakers, teachers...) to contribute to its development,

For this project, we chose to adopt as much as possible the development paradigm of LINUX and GNU software[2], as we believe that the lack of high level, rich and freely accessible multilingual lexical data is one of the most crucial obstacle for the development of a truly multilingual information society[3].

[1] http://www.papillon-dictionary.org/

[2] i.e. allowing and encouraging external users to *access and contribute to* the database.

[3] i.e. an Information Society with no linguistic domination and where everybody will be able to access any content in its own mother tongue.

2.2 Papillon acception based multilingual database

The Papillon multilingual database has been designed independently of its usage(s). It consists in several monolingual volumes linked by way of a single interlingual volume called the interlingual acception dictionary.

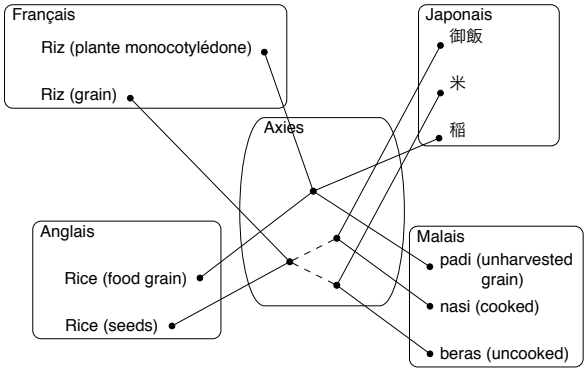


Figure 1: Macrostructure of the Papillon MLDB, showing the handling of contractive problems.

Each monolingual volume consists in a set of word senses (*lexies*), each lexie being described using a structure derived from the Explanatory and Combinatory Dictionary (Mel'čuk et al., 1995; Mel'čuk et al., 1984 1989 1995 1996).

The interlingual acception dictionary consists in a set of interlingual acceptions (*axies*) as defined in (Sérasset, 1994). An interlingual acception serves as a placeholder bearing links to lexies and links between axies[4]. This simple mechanism allows for the coding of translations. As an example, figure 1 shows how we can represent a quadrilingual database with contrastive problems (on the well known "rice" example).

2.3 Development methodology

The development of the Papillon multilingual dictionary gathers voluntary contributors and trusted language specialist involved in different tasks (as shown in figure 2).

- First, an automatic process creates a draft acception based multilingual lexical database from existing monolingual and bilingual lexical resources as shown in (Teeraparseree, 2003; Mangeot-Lerebours et al., 2003). This step is called the *bootstrapping* process.

[4]Note that these links are not interpreted semantically, but only reflect the fact that translation is possible

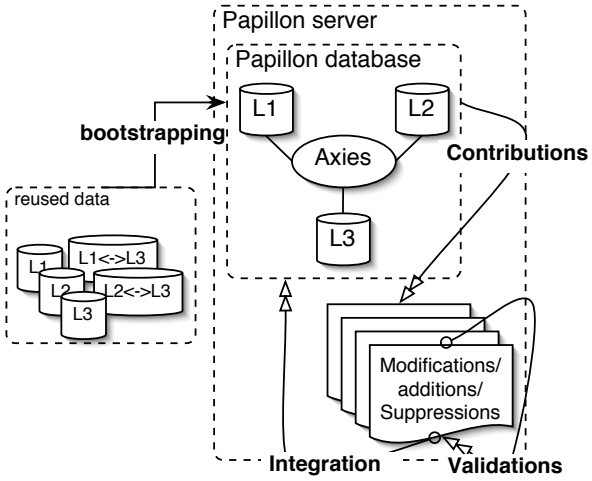


Figure 2: Methodology for the development of the Papillon database.

- Then, *contributions* may be performed by volunteers or trusted language specialists. A contribution is either the modification of an entry, its creation or its deletion. Each contribution is stored and immediately available to others.
- Volunteers or language specialist may *validate* these contributions by ranking them.
- Finally, trusted language specalists will *integrate* the contribution and apply them to the master MLDB. Rejected contributions won't be available anymore.

2.4 The Papillon Platform

The Papillon platform is a community web site specifically developed for this project. This platform is entirely written in Java using the "Enhydra[5]" web development Framework. All XML data is stored in a standard relational database (Postgres). This community web site proposes several services:

- a unified interface to simultaneously *access* the Papillon MLDB and several other monolingual and bilingual dictionaries;
- a specific edition interface to *contribute* to the Papillon MLDB,
- an open document repository where registered users may share writings related to the project; among these documents, one may find all the papers presented in the

[5]see http://www.enhydra.org/

different Papillon workshops organized each year by the project partners;

- a mailing list archive,

Sections 3 and 4 present the first and second services.

3 Unified access to existing dictionaries

3.1 Presentation

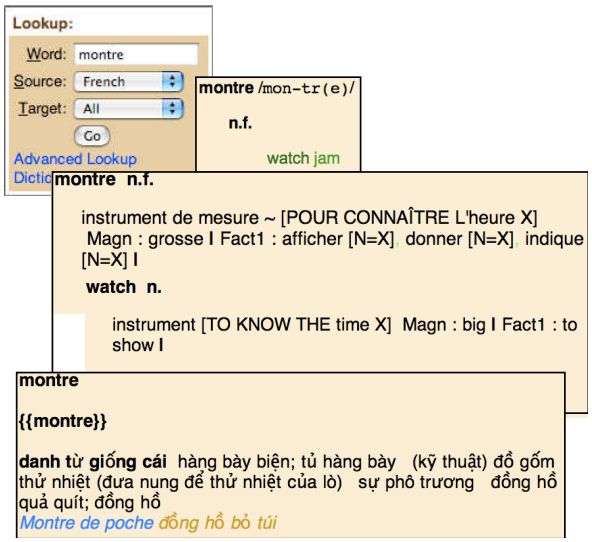


Figure 3: The unified access interface and results from three different dictionaries

To encourage volunteers, we think that it is important to give a real service to attract as many Internet users as possible. As a result, we began our development with a service to allow users to access to many dictionaries in a unified way. This service currently gives access to twelve (12) bilingual and monolingual dictionaries, totalizing a little less than 1 million entries, as detailled in table 1.

3.2 Strong points

The unified access interface allows the user to access simultaneously to several dictionaries with different structures. All available dictionary will be queried according to its own structure. Moreover, all results will be displayed in a form that fits its own structure.

Any monolingual, bilingual or multilingual dictionary may be added in this collection, provided that it is available in XML format.

With the Papillon platform, giving access to a new, unknown, dictionary is a matter of writing 2 XML files: a dictionary description and an XSL stylesheet. For currently available dictionaries, this took an average of about one hour per dictionary.

Dictionary	Languages	Nb of Entries
Armament[a]	fra eng	1116
Cedict[b]	zho eng	215424
Ding[c]	deu eng	124413
Engdict[d]	eng kor	214127
FeM[e]	fra eng msa	19247
Homerica[f]	fra	441
JMDict[g]	jp en fr de	96264
KanjiDict[h]	jpn eng	6355
Papillon	multi	1323
ThaiDict[i]	tha	10295
VietDict[j]	fra vie	41029
WaDokuJiTen[k]	jpn deu	214274

[a] Japanese French dictionary of armament from the French Embassy in Japan
[b] Chinese English from Mandel Shi (Xiamen univ.)
[c] (Richter, 1999)
[d] (Paik and Bond, 2003)
[e] (Gut et al., 1996)
[f] University Stendhal, Grenoble III
[g] (Breen, 2004a)
[h] (Breen, 2004b)
[i] Thai Dictionary of Kasetsart University
[j] (Duc, 1998)
[k] (Apel, 2004)

Table 1: Dictionaries available through the unified access interface

3.3 Implementation

It is possible to give access to any XML dictionary, regardless of its structure. For this, you have to identify a minimum set of information in the dictionary's XML structure.

The Papillon platform defines a standard structure of an abstract dictionary containing the most frequent subset of information found in most dictionaries. This abstract structure is called the Common Dictionary Markup (Mangeot-Lerebours and Sérasset, 2002). To describe a new dictionary, one has to write an XML file that associate CDM element to pointers in the original dictionary structure.

As an example, the French English Malay FeM dictionary (Gut et al., 1996) has a specific structure, illustrated by figure 4.

Figure 5 gives the XML code associating elements of the FeM dictionary with elements of the CDM.

Along with this description, one has to de-

```
<HFEM xmlns:xml="http://www.w3.org/.../namespace">
  <HW-FRE>montre</HW-FRE>
  <HOM/>
  <PRNC>mon-tr(e)</PRNC>
  <AUX/>
  <BODY>
    <SENSE-STAR>
      <SENSE>
        <CAT-STAR>n.f.</CAT-STAR>
        <SENSE1-STAR>
          <SENSE1>
            <TRANS-STAR>
              <TRANS>
                <ENG-STAR>watch</ENG-STAR>
                <MAL-STAR>jam</MAL-STAR>
              </TRANS>
            </TRANS-STAR>
            <EXPL-STAR/>
          </SENSE1>
        </SENSE1-STAR>
      </SENSE>
    </SENSE-STAR>
  </BODY>
</HFEM>
```

Figure 4: A simplified example entry from the French English Malay FeM dictionary.

```
<cdm-elements>
  <cdm-volume element="volume"/>
  <cdm-entry element="HFEM"/>
  <cdm-headword element="HW-FRE"/>
  <cdm-pronunciation element="PRNC"/>
  <cdm-pos element="CAT-STAR"/>
  <cdm-definition element="FRE"/>
  <cdm-translation d:lang="eng"
                   element="ENG-STAR"/>
  <cdm-translation d:lang="msa"
                   element="MAL-STAR"/>
  <cdm-example d:lang="fra" element="FRE"/>
  <cdm-example d:lang="eng" element="ENG"/>
  <cdm-example d:lang="msa" element="MAL"/>
  <cdm-key1 element="HOM"/>
</cdm-elements>
```

Figure 5: Associations between elements of the FeM dictionary and elements of the CDM.

fine an XSL style sheet that will be applied on requested dictionary elements to produce the HTML code that defines the final form of the result. If such a style sheet is not provided, the Papillon platform will itself transform the dictionary structure into a CDM structure (using the aforementioned description) and apply a generic style sheet on this structure.

4 Editing dictionaries entries

4.1 Presentation

As the main purpose of the Papillon platform is to gather a community around the *development* of a dictionary, we also developed a service for the edition of dictionary entries.

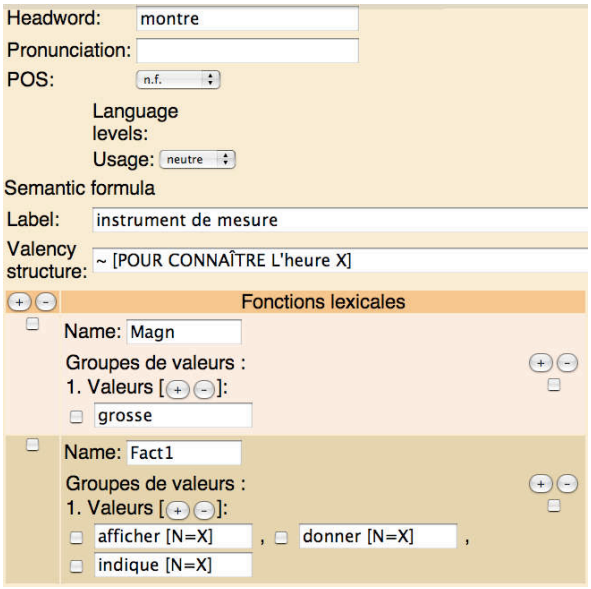


Figure 6: The edition interface is a standard HTML interface

Any user, who is registered and logged in to the Papillon web site, may contribute to the Papillon dictionary[6] by creating or editing[7] an entry. Moreover, when a user asks for an unknown word, he is encouraged to contribute it to the dictionary.

Contribution is made through a standard HTML interface (see figure 6). This interface is rather crude and raises several problems. For instance, there is no way to copy/paste part of an existing entry into the edition window. Moreover, editing has to be done on-line[8]. However, as the interface uses only standard HTML elements with minimal javascript functionality, it may be used with any Internet browser on any platform (provided that the browser/platform correctly handles unicode forms).

4.2 Strong points

From the beginning, we wanted this interface to be fully customizable by Papillon members

[6] And, for now, only to this particular dictionary.

[7] Removal of an entry is not yet implemented.

[8] In fact, entries may be edited off-line and uploaded on the server, but there is currently no specialized interface for off-line edition, meaning that users will have to use standard text/XML editor for this.

without relying on the availability of a computer science specialist. our reasons are:

- the fact that we wanted the structure of the Papillon dictionary to be adaptable along with the evolution of the project, without implying a full revisit of the web site implementation;
- the fact that each language may slightly adapt the Papillon structure to fit its own needs (specific set of part of speech, language levels, etc.), hence adding a new dictionary implies adding a new custom interface;

Hence, we chose to develop a system capable of generating a usable interface from a) a description of the dictionary structure (an XML Schema) and b) a description of the mapping between element of the XML structure and standard HTML inputs.

For this, we used the ARTStudio tool described by (Calvary et al., 2001). Using a tool that allows for the development of plastic user interfaces allows us to generate not only one, but several interfaces on different devices. Hence, as we are now able to generate an HTML interface usable with any standard web browser supporting Unicode, we may, in the future, generate interfaces for Java applications (that can be used offline) or interfaces for portable devices like pocket PCs or Palm computers.

4.3 Implementation

4.3.1 Definition of the dictionary structure

To provide an edition interface, the Papillon platform needs to know the exact dictionary structure. The structure has to be defined as a standard XML schema. We chose to use XML schema because it allows for a finer description compared to DTDs (for instance, we may define the set of valid values of the textual content of an XML element). Moreover XML schemata provides a simple inheritance mechanism that is useful for the definition of a dictionary. For instance, we defined a general structure for the Papillon dictionary (figure 7) and used the inheritance mechanism to refine this general structure for each language (as in figure 8).

4.3.2 Description of the interface

Describing the interface is currently the most delicate required operation. The first step is to define the set of elements that will appear in the interface and their relation with the dictionary structure. Each such element is given a unique ID. This step defines an abstract interface where all elements are known, but not their layout, nor their kind.

This step allows for the definition of several different tasks for the edition of a single dictionary.

The second step is to define the concrete realization and the position of all these elements.

```
<element name="lexie">
 <complexType>
  <sequence>
   <element ref="d:headword" minOccurs="1"
                             macOccurs="1" />
   <element ref="d:writing" ... />
   <element ref="d:reading" ... />
   <element ref="d:pronunciation" ... />
   <element ref="d:pos" ... />
   <element ref="d:language-levels" ... />
   <element ref="d:semantic-formula" ... />
   <element ref="d:government-pattern" .../>
   <element ref="d:lexical-functions" ... />
   <element ref="d:examples" ... />
   <element ref="d:full-idioms" ... />
   <element ref="d:more-info" ... />
  </sequence>
  <attribute ref="d:id" use="required" />
 </complexType>
</element>
...
<element name="pos" type="d:posType" />
<simpleType name="posType">
  <restriction base="string" />
</simpleType>
...
```

Figure 7: General structure shared by all volumes of the Papillon dictionary; showing the part of speech element `pos` defined as a textual element.

```
<simpleType name="posType">
  <restriction base="d:posType">
    <enumeration value="n.m." />
    <enumeration value="n.m. inv." />
    <enumeration value="n.m. pl." />
    <enumeration value="n.m., f." />
    <enumeration value="n.f." />
    <enumeration value="n.f. pl." />
    ...
  </restriction>
</simpleType>
```

Figure 8: Redefinition of the type of the part of speech `pos` element in the Papillon French definition.

For instance, in this step, we specify the POS element to be rendered as a menu. Several kind of widgets are defined by ARTStudio. Among them, we find simple HTML inputs like text boxes, menus, check-boxs, radio buttons, labels..., but we also find several high level elements like generic lists of complex elements.

As an simple example, we will see how the `pos` (part of speech) element is rendered in the Papillon interface. First, there will be an interface element (called S.364) related to the `pos` element (figure 9). Second, this element will be realized in our interface as a comboBox (figure 10).

```
<Instance type="element" id="S.364">
 <InstanceKind value="static"/>
 <InstanceBuildKind value="regular"/>
 <Name value="pos"/>
 <ClassNameSpace value=""/>
 <ClassName value="posType"/>
 <TaskOwnerID value="S.360"/>
 <TaskRangeID list="S.360"/>
</Instance>
```

Figure 9: Definition of the abstract interface element associated to the `pos` element. This element will display/edit value of type `posType` defined in the aforementioned schema.

```
<Interactor type="element"
            class="GraphicInteractor" id="i2008">
 <Type value="presentation"/>
 <TaskID value="S.363"/>
 <InteractorID value="ComboBox"/>
 <InstanceID value="S.364"/>
 <Width value="10"/>
 <Height value="20"/>
</Interactor>
```

Figure 10: Definition of the effective widget for the `pos` element.

Using this technique is rather tricky as there is currently no simple interface to generate these rather complex descriptions. However, using these separate description allows the definition of several edition tasks (depending on the user profile) and also allows, for a single task, to generate several concrete interfaces, depending on the device that will be used for edition (size of the screen, methods of interactions, etc.).

4.3.3 Interface generation

Using the describe structure of the dictionary, we are able to generate an empty dictionary entry containing all mandatory elements. Then, we walk this structure and instantiate all associated widgets (in our case HTML input elements), as defined in the interface description. This way, we are able to generate the corresponding HTML form.

When the user validates a modification, values of the HTML input elements are associated to the corresponding parts of the edited dictionary structure (this is also the case if the user asks for the addition/suppression of an element in the structure). Then, we are able to regenerate the interface for the modified structure. We iterate this step until the user saves the modified structure.

5 Conclusions

The Papillon platform is still under development. However, it already proves useful for the *diffusion* of a little less than 1 million entries from 12 very different dictionaries. This is possible as, from the very beginning, we designed the platform to be as a generic as possible.

This genericity also allows for its use for the *on-line development* of the Papillon database. It is also used for the development of the Estonian French GDEF dictionary, managed by Antoine Chalvin from INALCO, Paris. Moreover, we developed an interface for the japanese German WadokujiTen (Apel, 2004). This proves that our platform may be useful in a general context.

Our future activities will follow 3 axis:

- improving the definition of edition interfaces; currently, we have no tool to simplify this definition and its complexity makes it difficult for a linguist to use it without help from computer science specialists;
- generating different interfaces from the same descriptions; currently, we only generate on-line HTML interfaces, but the tools we use allows for the development of interfaces in other contexts; hence with the same approach, we will develop java applets or java applications to be used either on-line or off-line;
- developing network cooperation modules between several instances of the Papillon platform; this will allow the deployment of the platform on several sites; we will address two aspects of such a deployment; first, duplication of identical instances providing access and edition services on the same dictionaries; second the deployment

of several instances providing access and edition services on different dictionaries (where dictionaries edited on a site may be accessed on another site).

6 Acknowledgements

Developments on the Papillon project could not have taken place without support from CNRS (France) and JSPS (Japan). We would like to warmly thank François Brown de Colstoun who supports this project since its very beginning. Developments of the platform and especially the editing part has been mainly done by Mathieu Mangeot and David Thevenin during their Post Doctoral fellowship at NII (National Institue of Informatics), Tokyo. Finally the Papillon platform would not be useful without partners who agreed to give free access to their superb dictionaries.

References

Ulrich Apel. 2004. WaDokuJT - A Japanese-German Dictionary Database. In *Papillon 2002 Workshop on Multilingual Lexical Databases*, NII, Tokyo, Japan, 6-18 July.

Jim W. Breen. 2004a. JMdict: a Japanese-Multilingual Dictionary. In Gilles Sérasset, Susan Armstrong, Christian Boitet, Andrei Pospescu-Belis, and Dan Tufis, editors, *post COLING Wokshop on Multilingual Linguistic Resources*, Geneva, Switzerland, 28th august. International Committee on Computational Linguistics.

Jim W. Breen. 2004b. Multiple Indexing in an Electronic Kanji Dictionary. In Michael Zock and Patrick St Dizier, editors, *post COLING workshop on Enhancing and Using Electronic Dictionaries*, Geneva, Switzerland, 29th august. International Committee on Computational Linguistics.

Gaëlle Calvary, Joëlle Coutaz, and David Thevenin. 2001. A unifying reference framework for the development of plastic user interfaces. In M. Reed Little and L. Nigay, editors, *Engineering for Human-Computer Interaction: 8th IFIP International Conference, EHCI 2001*, volume 2254 / 2001 of *Lecture Notes in Computer Science*, page 173. Springer-Verlag Heidelberg, Toronto, Canada, May.

Ho Ngoc Duc, 1998. *Vietnamese French Online Dictionary.* http://www.informatik.uni-leipzig.de/~duc/Dict/.

Yvan Gut, Puteri Rashida Megat Ramli, Zaharin Yusoff, Kim Choy Chuah, Salina A. Samat, Christian Boitet, Nicolai Nedobejkine, Mathieu Lafourcade, Jean Gaschler, and Dorian Levenbach. 1996. *Kamus Perancis-Melayu Dewan, Dictionnaire francais-malais.* Dewan Bahasa dan Pustaka, Kuala Lumpur.

Mathieu Mangeot-Lerebours and Gilles Sérasset. 2002. Frameworks, implementation and open problems for the collaborative building of a multilingual lexical database. In Grace Ngai, Pascale Fung, and Kenneth W. Church, editors, *Proc. of SEMANET Workshop, Post COLING 2002 Workshop*, pages 9–15, Taipei, Taiwan, 31 August.

Mathieu Mangeot-Lerebours, Gilles Sérasset, and Mathieu Lafourcade. 2003. Construction collaborative dune base lexicale multilingue, le projet Papillon. *TAL*, 44(2):151–176.

Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Eltnisky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, Adèle Lessard, Alain Polguère, and Suzanne Mantha. 1984, 1989, 1995, 1996. *Dictionnaire Explicatif et Combinatoire du français contemporain, recherches lexico-sémantiques, volumes I, II, III et IV.* Presses de l'Université de Montréal, Montréal(Quebec), Canada.

Igor Mel'čuk, Andre Clas, and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire.* Universites francophones et champs linguistiques. AUPELF-UREF et Duculot, Louvain-la Neuve.

Kyonghee Paik and Francis Bond. 2003. Enhancing an English/Korean Dictionary. In *Papillon 2003 Workshop on Multilingual Lexical Databases*, Sapporo, Japan, 3-5 July.

Franck Richter, 1999. *Ding: a Dictionary Lookup Program.* http://www-user.tu-chemnitz.de/~fri/ding/.

Gilles Sérasset and Mathieu Mangeot-Lerebours. 2001. Papillon lexical database project: Monolingual dictionaries and interlingual links. In *NLPRS-2001*, pages 119–125, Tokyo, 27-30 November.

Gilles Sérasset. 1994. Interlingual lexical organisation for multilingual lexical databases in nadia. In Makoto Nagao, editor, *COLING-94*, volume 1, pages 278–282, August.

Aree Teeraparseree. 2003. Jeminie: A flexible system for the automatic creation of interlingual databases. In *Papillon 2003 Workshop on Multilingual Lexical Databases*, Sapporo, Japan, 3-5 July.

Semi-Automatic Construction of Korean-Chinese Verb Patterns Based on Translation Equivalency

Munpyo Hong
Dept. of Speech/Language Technology Research, ETRI
Korea
Hmp63108@etri.re.kr

Young-Kil Kim
Dept. of Speech/Language Technology Research, ETRI
Korea
kimyk@etri.re.kr

Sang-Kyu Park
Dept. of Speech/Language Technology Research, ETRI
Korea
parksk@etri.re.kr

Young-Jik Lee
Dept. of Speech/Language Technology Research, ETRI
Korea
ylee@etri.re.kr

Abstract

This paper addresses a new method of constructing Korean-Chinese verb patterns from existing patterns. A verb pattern is a subcategorization frame of a predicate extended by translation information. Korean-Chinese verb patterns are invaluable linguistic resources that are not only used for Korean-Chinese transfer but also for Korean parsing. Usually a verb pattern has been either hand-coded by expert lexicographers or extracted automatically from bilingual corpus. In the first case, the dependence on the linguistic intuition of lexicographers may lead to the incompleteness and the inconsistency of a dictionary. In the second case, extracted patterns can be domain-dependent. In this paper, we present a method to construct Korean-Chinese verb patterns semi-automatically from existing Korean-Chinese verb patterns that are manually written by lexicographers.

1 Introduction

PBMT (Pattern-based Machine Translation) approach has been adopted by many MT researchers, mainly due to the portability, customizability and the scalability of the approach. cf. Hong et al. (2003a), Takeda (1996), Watanabe & Takeda (1998). However, major drawback of the approach is that it is often very costly and time-consuming to construct a large amount of data enough to assure the performance of the PBMT system. From this reason many studies from PBMT research circles have been focused on the data acquisition issue. Most of the data acquisition studies were about automatic acquisition of lexical resources from bilingual corpus.

Since 2001, ETRI has developed a Korean-Chinese MT system, TELLUS K-C, under the auspices of the MIC (Ministry of Information and Communication) of Korean government. We have adopted verb pattern based approach for Korean-Chinese MT. The verb patterns play the most crucial role not only in the transfer but also in the source language analysis. In the beginning phase of the development, most of the verb patterns were constructed manually by experienced Korean-Chinese lexicographers with some help of editing tools and electronic dictionaries. In the setup stage of a system, the electronic dictionary is very useful for building a verb pattern DB. It provides with a comprehensive list of entries along with some basic examples to be added to the DB. In most cases, however, the examples in the dictionary with which the lexicographers write a verb pattern are basic usages of the verb in question, and other various usages of the verb are often neglected. Bilingual corpus can be useful

resources to extract verb patterns. However, as for language pairs like Korean-Chinese for which there are not so much bilingual corpus available in electronic form, the approach does not seem to be suitable. Another serious problem with the bilingual corpus-based approach is that the patterns extracted from the corpus can be domain-dependent.

The verb pattern generation based on translation equivalency is another good alternative to data acquisition from bilingual corpus. The idea was originally introduced by Fujita & Bond (2002) for Japanese to English MT.

In this paper, we present a method to construct Korean-Chinese verb patterns from existing Korean-Chinese verb patterns that are manually written by lexicographers. The clue for the semi-automatic generation is provided by the idea that verbs of similar meanings often share the argument structure as already shown in Levin (1993). The synonymy among Korean verbs can be indirectly inferred from the fact that they have the same Chinese translation.

We have already applied the approach to TELLUS K-C and increased the number of verb patterns from about 110,000 to 350,000. Though 350,000 patterns still contain many erroneous patterns, the evaluations in section 5 will show that the accuracy of the semi-automatically generated patterns is noteworthy and the pattern matching ratio improves significantly with 350,000 pattern DB.

2 Related Works

When constructing verb pattern dictionary, too much dependence on the linguistic intuition of lexicographers can lead to the inconsistency and the incompleteness of the pattern dictionary. Similar problems are encountered when working with a paper dictionary due to the insufficient examples. Hong et al (2002) introduced the concept of causative/passive linking to Korean word dictionary. The active form 'mekta (to eat)' is linked to its causative/passive forms 'mekita (to let eat)', and 'mekhita (to be eaten)', respectively. The linking information of this sort helps lexicographers not to forget to construct verb patterns for causative/passive verbs when they write a verb pattern for active verbs. The semi-automatic generation of verb patterns using translation equivalency was tried in Hong et al (2002). However, as only the voice information was used as a filter, the over-generation problem is serious.

Fujita & Bond (2002) and Bond & Fujita (2003) introduced the new method of constructing a new valency entry from existing entries for Japanese-English MT. Their method creates valency patterns for words in the word dictionary whose English translations can be found in the valency dictionary. The created valency patterns are paraphrased using monolingual corpus. The human translators check the grammaticality of the paraphrases.

Yang et al. (2002) used passive/causative alternation relation for semi-automatic verb pattern generation. Similar works have been done for Japanese by Baldwin & Tanaka (2000) and Baldwin & Bond (2002) .

3 Verb Pattern in TELLUS K-C

The term 'verb pattern' is understood as a kind of subcategorization frame of a predicate. However, a verb pattern in our approach is slightly different from a subcategorization frame in the traditional linguistics. The main difference between the verb pattern and the subcategorization frame is that a verb pattern is always linked to the target language word (the predicate of the target language). Therefore, a verb pattern is employed not only in the analysis but also in the transfer phase so that the accurate analysis can directly lead to the natural and correct generation. In the theoretical linguistics, a subcategorization frame always contains arguments of a predicate. An adjunct of a predicate or a modifier of an argument is usually not included in it. However, in some cases, these words must be taken into account for the proper translation. In translations adjuncts of a verb or modifiers of an argument can seriously affect the selection of target words. (1) exemplifies verb patterns of "cata (to sleep)":

(1)
cata1 : A=WEATHER!ka ca!ta[1] > A 停:v
[param(A)ka cata: *The wind has died down*]

[1] The slot for nominal arguments is separated by a symbol "!" from case markers like "ka", "lul", "eykey", and etc. The verb is also separated by the symbol into the root and the ending.

cata2 : A=HUMAN!ka ca!ta > A 睡觉:v
[ai(A)ka cata: *A baby is sleeping*]
cata 3 : A=WATCH! ka ca!ta > A 停:v
[sikye(A)ka cata: *A watch has run down*]
cata 4 : A=PHENOMENA!ka ca!ta > A 平静:v
[phokpwungwu(A)ka cata: *The storm has abated*]

On the left hand of ">" Korean subcategorization frame is represented. The argument position is filled with a variable (A, B, or C) equated with a semantic feature (WEATHER, HUMAN, WATCH, PHENOMENA). Currently we employ about 410 semantic features for nominal semantic classifications. The Korean parts of verb patterns are employed for syntactic parsing.

On the right hand of ">" Chinese translation is given with a marker ":v". To every pattern is attached an example sentence for better comprehensibility of the pattern. This part serves for the transfer and the generation of Chinese sentence.

4 Pattern Construction based on Chinese Translation

In this chapter, we elaborate on the method of semi-automatic construction of Korean-Chinese verb patterns. Our method is similar to that of Fujita & Bond (2002) and inspired by it as well, i.e. it makes most use of the existing resources.

The existing resources are in this case verb patterns that have already been built manually. As every Korean verb pattern is provided with the corresponding Chinese translation, Korean verb patterns can be re-sorted to Chinese translations. The basic assumption of this approach is that the verbs with similar meanings tend to have similar case frames, as is pointed out in Levin (1993). As an indication to the similarity of meaning among Korean verbs, Chinese translation can be employed. If two verbs share Chinese translation, they are likely to have similar meanings. The patterns that have translation equivalents are seed patterns for automatic pattern generation.

Our semi-automatic verb pattern generation method consists of the following four steps:

Step1: Re-sort the existing Korean-Chinese verb patterns according to Chinese verbs

Example:

Chinese Verb 1: 给 (to give)

tulita	A=HUMAN!ka B=CAR!lul tuli!ta
cwuta	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul cwu!ta
swuyehata	A=HUMAN!ka B=MONEY!lul swuyeha!ta

Chinese Verb 2: 停止 (to stop)

kumantwuta	A=HUMAN!ka B=CONSTRUCTION!lul kumantwu!ta
kwantwuta	A=ORGANIZATION!ka B=VIOLATION!lul kumantwu!ta

When the re-sorting is done, we have sets of synonymous Korean verbs which share Chinese translations, such as {tulita, cwuta, swuyehata} and {kumantwuta, kwantwuta }.

Step2: Pair verbs with the same Chinese translation

Example:

Chinese Verb 1: 给 (to give)

Pair1:

tulita	A=HUMAN!ka B=CAR!lul tuli!ta
cwuta	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul cwu!ta

Pair2:

tulita	A=HUMAN!ka B=CAR!lul tuli!ta
swuyehata	A=HUMAN!ka B=MONEY!lul swuyeha!ta

Pair3:

cwuta	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul cwu!ta
swuyehata	A=HUMAN!ka B=MONEY!lul

	swuyeha!ta

Step3: Exchange the verbs, if the following three conditions are met:

- The two Korean verbs of the pair have the same voice information
- Neither of the two verbs is idiomatic expressions
- The Chinese translation is not 加以, 进行, 做, 作

Example:

tulita	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul tuli!ta
tulita	A=HUMAN!ka B=MONEY!lul tuli!ta
cwuta	A=HUMAN!ka B=CAR!lul cwu!ta
cwuta	A=HUMAN!ka B=MONEY!lul cwu!ta
swuyehata	A=HUMAN!ka B=CAR!lul swuyeha!ta
swuyehata	A=HUMAN!ka B=HUMAN!eykey C=VEGETABLE!lul swuyeha!ta

Step4: If the newly-generated pattern already exists in the verb pattern dictionary, it is discarded.

The three conditions to be met in the third step are the filters to prevent the over-generation of patterns. The following examples shows why the first condition, i.e., "the voice of the verbs in question must agree", must be met.

(2) 漂 (to float)

ttuta : A=PLANT!ka B=PLACE!ey ttu!ta > A 漂:v 在 B 上 [namwutip(A)i mwulwi(B)ey ttuta: *A leaf is floating on the water*]

ttiwuta : A=HUMAN!ka B=PLACE!ey C=PLANT!lul ttiwu!ta > A 使 C 漂:v 在 B 上 [ai(A)ka mwulwi(B)ey namwutip(C)ul ttiwuta: A baby floated a leaf on the water]

(3) 滥用 (to use)

sayongtoyta : A=HUMAN!eyuyhay B=MEDICINE!ka sayongtoy!ta > B 被 A 滥用:v [hankwuksalamtul(A)eyuyhay yak(B)i hambwulo sayongtoyta: *The drug is misused by Koreans*]

sayonghata : A=HUMAN!ka B=MEDICINE!lul sayongha!ta > A 滥用:v B [hankwuksalamtul (A)un yak(B)ul hambwulo sayonghanta: *Koreans are misusing the drug*]

As we re-sort the existing patterns according to the Chinese verbs which are marked with ":v", the verbs of different voice may be gathered together. However, as the above examples show, the voice (active vs. causative in (2), passive vs. active in (3)) affects the argument structure of verbs. We conclude that generating patterns without considering the voice information can lead to the over-generation of patterns. The voice information of verbs can be obtained from the linking information between the verb pattern dictionary and the word dictionary. We will not look into the details of the linking relation between the verb pattern dictionary and the word dictionary of TELLUS K-C system in this paper. cf. Hong et al. (2002)

The second condition relates to the lexical patterns of Korean. Lexical patterns are used for collocational expressions. As the nature of collocation implies, a predicate that shows a strict co-occurrence relation with a certain nominal argument cannot be arbitrarily combined with any other nouns.

The third condition deals with the support verb construction of Chinese. The four verbs, 加以, 进行, 做, 作, belong to the major verbs in Chinese that form support verb construction with predicative nouns. In support verb construction, the argument structure of the sentence is not determined by a verb but by a predicative noun. Because of this, the same Chinese translation cannot be the indication of similar meaning of Korean verbs, as followed:

(4) 作:v (to make)
ttallangkelita (to ring): A=BELL!ka
ttallangkeli!ta > A 作:v 厩当声
[pangwul(A)i ttallangkelita: *A bell is ringing*]

ssawuta1 (to fight) : A=HUMAN!ka B=PROPERTY!wa ssawu!ta > A 为 B 作:v 斗争 [kunye(A)ka mwulka(B)wa ssawunta: *She is struggling with high price*]

wuntonghata (to exercise) : A=HUMAN!ka B=PLACE!eyse wuntongha!ta > A 在 B 作:v 运动 [ku(A)ka chewyukkwan(B)eyse wuntonghanta: *He is exercising in the gymnasium*]

Although the Korean verbs "ttallangkelita (to ring)", "ssawuta (to fight)", "wuntonghata (to exercise)" share the Chinese verb "作", the argument structure of each Chinese translation is determined by the predicative nouns that are syntactically objects of the verbs.

5 Evaluation

The 114,581 verb patterns we have constructed for 3 years were used as seed patterns for semi automatic generation of patterns. After the steps 1 and 2 of the generation process were finished, the sets of possible synonymous verbs were constructed. To filter out the wrong synonym sets, the whole sets were examined by two lexicographers. It took a week for two lexicographers to complete this process. The wrong synonym sets were produced mainly due to the homonymy of Chinese verbs.

From the original 114,581 patterns, we generated 235,975 patterns. We performed two evaluations with the generated patterns. In the first evaluation, we were interested in finding out how many correct patterns were generated. The second evaluation dealt with the improvement of the pattern matching ratio due to the increased number of patterns.

Evaluation 1

In the first evaluation we randomly selected 3,086 patterns that were generated from 30 Chinese verbs. The expert Korean-Chinese lexicographers examined the generated patterns. Among the 3,086 patterns, 2,180 were correct. The accuracy of the semi-automatic generation was 70.65%. Although the evaluation set was relatively small in size, the accuracy rate seemed to be quite promising, considering there still remain other filtering factors that can be taken into account additionally.

Chinese Verbs	30
Unique generated patterns	3,086
Correct patterns	2,180
Erroneous patterns	906
Accuracy	70.65%

Table 1: Accuracy Evaluation

The majority of the erroneous patterns can be classified into the following two error types:

- The verbs share similar meanings and selectional restrictions on the arguments. However, they differ in selecting the case markers for argument positions (the most prominent error).

  Ex) **~eykey** masseta/ **~wa** taykyelhata (to face somebody)

- The verbs share similar meanings, but the selectional restrictions are different.

  Ex) **PAPER**!lul kyopwuhata (to deliver) / **MONEY**!lul nappwuhata (to pay)

Evaluation 2

In the second evaluation, our interest was to find out how much improvement of pattern matching ratio can be achieved with the increased number of patterns in comparison to the original pattern DB. For the evaluation, 300 sentences were randomly extracted from various Korean newspapers. The test sentences were about politics, economics, science and sports. In the 300 sentences there were 663 predicates.

With the original verb pattern DB, i.e. with 114,581 patterns, the perfect pattern matching ratio was 59.21%, whereas the perfect matching ratio rose to 64.40% with the generated pattern DB.

	114,581 Verb patterns	**350,556 Verb patterns**

Num. Of Sentences	300	
Num. of. Predicates	663	
Perfect Matching	392	427
No Matching	73	66
Perfect Matching Ratio	59.21 %	64.40 %

Table 2: Pattern Matching Ratio Evaluation

6 Conclusion

Korean-Chinese verb patterns are invaluable linguistic resources that cannot only be used for Korean-Chinese transfer but also for Korean analysis. In the set-up stage of the development, a paper dictionary can be used for exhaustive listing of entry words and the basic usages of the words. However, as the verb patterns made from the examples of a dictionary are often insufficient, a PBMT system suffers from the coverage problem of the verb pattern dictionary. Considering there are not so many Korean-Chinese bilingual corpus available in electronic form till now, we believe the translation-based approach, i.e. Chinese-based pattern generation approach provides us with a good alternative.

The focus of our future research will be given on the pre-filtering options to prevent over-generation more effectively. Another issue will be about post-filtering technique using monolingual corpus with minimized human intervention.

References

T. Baldwin and F. Bond. 2002. Alternation-based Lexicon Reconstruction, *TMI 2002*

T. Baldwin and H. Tanaka. 2000. Verb Alternations and Japanese – How, What and Where? *PACLIC2000*

F. Bond and S. Fujita. 2003. Evaluation of a Method of Creating New Valency Entries, *MT-Summit 2002*

S. Fujita and F. Bond. 2002. A Method of Adding New Entries to a Valency Dictionary by Exploiting Existing Lexical Resources, *TMI2002*

M. Hong, Y. Kim, C. Ryu, S. Choi and S. Park. 2002. Extension and Management of Verb Phrase Patterns based on Lexicon Reconstruction and Target Word Information, *The 14th Hangul and Korean Language Processing* (in Korean)

M. Hong, K. Lee, Y. Roh, S. Choi and S. Park. 2003. Sentence-Pattern based MT revisited, *ICCPOL 2003*

B. Levin. 1993. English verb classes and alternation , The University of Chicago Press

K. Takeda. 1996. Pattern-based Machine Translation, *COLING 1996*

H. Watanabe and K. Takeda. 1998. A Pattern-based Machine Translation System Extended by Example-based Processing, *ACL 1998*

S. Yang, M. Hong, Y. Kim, C. Kim, Y. Seo and S. Choi. 2002. An Application of Verb-Phrase Patterns to Causative/Passive Clause, *IASTED 2002*

Bilingual Sign Language Dictionary for Learning a Second Sign Language without Learning the Target Spoken Language

Emiko SUZUKI, Mariko HORIKOSHI, Kyoko KAKIHANA
Information Processing Department.
Tokyo Kasei Gakuin Tsukuba Womens University
Azuma 3-1, Tsukuba
Ibaraki Japan 305-0031
{emiko, horikosi, kakihana}@cs.kasei.ac.jp

Abstract

This paper describes a bilingual sign language dictionary (Japanese Sign Language and American Sign Language) that can help people learn each sign language directly from their mother sign language. Our discussion covers two main points. The first describes the necessity for a bilingual dictionary. Since there is no "universal sign language", or real "international sign language," deaf people would need to learn at least four languages if they want to talk to people whose mother tongue is different from their own: their mother sign language, their mother spoken language (as an intermediate language), the target spoken language, and the sign language fot the language in which they wish to communicate. The two spoken languages become language barriers for deaf people, and our bilingual dictionary will remove these barriers. The second describes the use of computers. As the use of computers becomes more widespread, it has become more convenient to study using computer software and/or the Internet facilities. Our dictionary system provides deaf people with an easy means of access using their mother-sign language so that they don't have to overcome the barrier of learning the target-spoken language. It also provides a way for people who are going to learn two sign languages to look up new vocabulary. Further, we plan to examine how our dictionary system could be used to educate and assist deaf people.

1 Introduction

Nowadays many deaf people have an opportunity to study abroad and to lcarn a forcign language. But there are three barriers they must overcome to acquire the target sign language: the first barrier is their mother-spoken language; the second is the target-spoken language; and, the last is the target-sign language.

Generally, deaf people are bilingual since they have to learn their mother-spoken language and its sign language. In the United States, many universities offer American Sign Language(ASL) as a second foreign language. It is recognized as an independent languages in the U.S. In Japan, Japanese Sign Language(JSL) has not yet been recognized as an independent language.

One of the main purposes of our dictionary is to remove these language barriers and help deaf people improve their sign language abilityies based on their spoken language(Japanese or English)[Suzuki E., and Kakihana K. 2002].

2 American Sign Language (ASL) & Japanese Sign Language (JSL)

2.1 American Sign Language (ASL)

American Sign Language (ASL) is a complex visual–spatial language, used by the deaf community in the United States and English-speaking parts of Canada (Nakamura (1)). The number of ASL users is almost five-hundred thousand. It is the native language of many hearing-impaired people, as well as some hearing children born into deaf families. The ASL is derived from the Native American sign language, with some words taken from French sign language.

ASL shares no grammatical similarities to English and should not be considered in any way to be a broken, mimed, or gestural form of English. In terms of syntax, for example, ASL uses topic-comment syntax, while English uses subject-verb-object.

2.2 Japanese Sign Language (JSL)

There are two main sign languages in Japan: "Japanese sign language," and "Japanese oral sign language." The former is used by deaf people, and the latter is mainly used by volunteers. It is a

type of pidgin-signed Japanese, often used in formal situations, lectures, and speeches. The main difference between the two is the sequence of the words. Japanese sign language syntax is like spoken English, using subject-verb-object, whereas Japanese oral sign language syntax is like spoken Japanese, using subject-object-verb. In this paper, since we will only discuss the sign language word dictionary and not with syntax, we will use the acronym "JSL" to refer to both Japanese sign languages.

2.3 Language Selection

As previously mentioned in terms of syntax, ASL has more in common with spoken Japanese than with English. For example, in spoken English, they say "What is your name?" and ASL signs "name"+"what". This word order is identical to oral Japanese. On the other hand, in JSL, the word order is "what"+"name" which is more like spoken English. That is one of the main reasons for us to focus on ASL as a bilingual dictionary. Another reason is that ASL is the fourth most commonly used language in the U.S.A. We assume that it is easier than learning another sign language for those who already know Japanese sign language (JSL) and are planning to learn a second sing language(Nakamura (2)). Further, according to some TV programs and newspaper reports, JSL is becoming more popular among Japanese recently. Therefore we decided to provide a bilingual dictionary for those who wish to learn JSL and ASL.

3 Digital sign language dictionary

3.1 The problem of digital sign language

Recently many digital dictionaries are available on the Internet or on CD-ROM. Some of the electronically accessible bilingual dictionaries and corpora include: English-French, German-English, Albanian-Spanish, English-Romanian, Greek-Russian, English-Spanish, English-Russian, English-Estonian, English-Hungarian, and Esperanto-English. These on-line dictionaries are easy to access by searching an Internet dictionary site.

Almost all of there digital dictionaries illustrate signs using cartoon-animations and not human gestures. We tried to make two types of dictionaries, and decided that the animations were more difficult to understand than human gestures especially by beginners. On the other hand, people who have been learning the sign language for a long time might be able to understand animated sign languages.

3.2 The purpose of our dictionary

Almost all of these digital dictionaries are for the people who can read and write their mother language fluently and not for those who have a disability with their mother tongue. The mother tongue for those who were born deaf is sign language, especially for those born into a deaf family. On the other hand, the main language for those who were born deaf but whose parents are not deaf is the parents' spoken language. The problem for these deaf children is that it is difficult for them to learn their spoken language in their country. Since their parents use their spoken language and have become used to using sign language for their child, a child who is deaf must learn at least two languages. When they want to or have to learn a foreign language, the foreign language becomes the second foreign language. The target sign language then, becomes their third one.

Sign language is believed to help those who want to communicate with people who have another mother tongue. We think that the second sign language helps deaf people to communicate with each other without learning the target-spoken language.

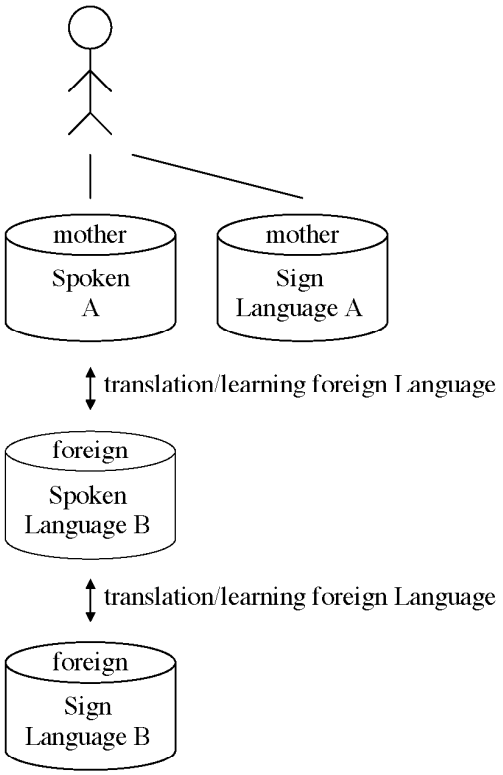


Figure1.Learning Sign Language Flow

We believe that our digital dictionary will remove the barriers for the deaf to learn another

foreign language. The users can choose the language (Japanese or English) when they want to search for signed words. They select the word from their favourite language and just look for it in the index. Our dictionary shows both Japanese and American sign languages with moving pictures taken from three directions.

4 Dictionary Configuration

4.1 Overview

Figure 2 shows our bilingual dictionary configuration. As you can see, we can search each sign language using either Japanese or English words. Thus, we can describe our system as a quadra-lingual dictionary system. Once you choose a word, our dictionary will show you the corresponding Japanese and American sign languages.

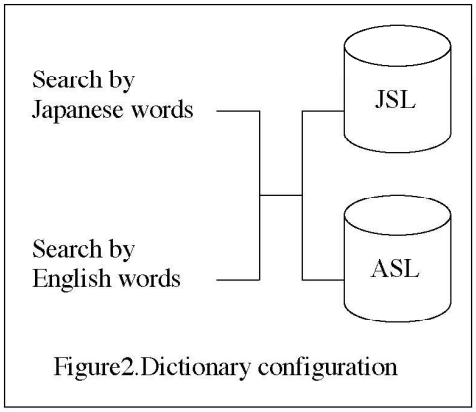


Figure2.Dictionary configuration

4.2 Search Flow

When this dictionary system starts up, the menu displays the languages the user can select (Fig. 3). As shown in Fig. 3, users can select "Exercise" from the menu after first learning some signs. Upon the selecting the language, the first characters of the indices are shown in a conventional alphabetical arrangement (Fig. 4).

Figure3. System Menu

Figure 4. First Characters of Japanaese Indices

For example, if Japanese is selected, the first characters of the indices are arranged in dictionary order (Fig. 4). While using Japanese to search the dictionary, the user can check the equivalent English words and ASL. Thus, they can learn spoken English and ASL simultaneously as shown in Fig. 5.

Figure 5. Dictionary Screen

The users can return to the Japanese indices and English indices just by clicking a button on the display using their mouse, as shown in Fig. 5. An example of the English indices that appeared on the screen is shown in Fig. 6. The resulting screen is how the screen looks like after the user clicks "w" for the search word in ASL. Also in Fig.6, you can see that English words are displayed befpre Japanese words.

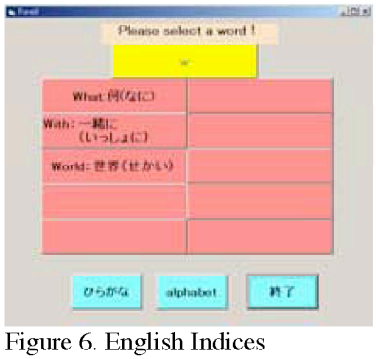


Figure 6. English Indices

4.3 Applications

We have improved the previous dictionary by adding an explanation of each sign operation and providing an example sentence for each word to enable natural language learnng.

As shown in Figure 7, eight more buttons have been added to the bilingual sign language display. Each are explained in Japanese and English and the example sentences are in Japanese and English.

Figure 7. Improved Dictionary Screen

When you press the "Explanation in Japanese" button, a new screen appears as shown in Figure 8.

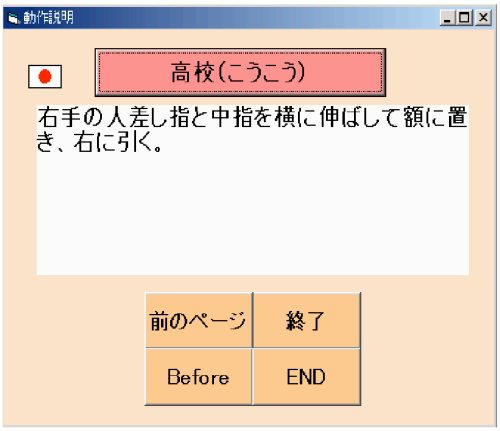


Figure 8. Explanation of signing in Japanese

5 Conclusion and future work

We have already completed a Japanese sign language dictionary with 50 entries that came from the JSL dictionary for beginners [Yonaiyama and Ogata 2001]. We also used corresponding American sign language entries, but some of the words do not exist. We are planning to add more Japanese and American sign language motion pictures. The cross-referenced features in our dictionary offer students, sign language learners, and deaf people, a genuine two-language resource that enhances the opportunity to obtain communication skills in both modes. We know that introducing ASL in the English learning classroom attracts the students' interest and is effective in learning a foreign language[Pauly, M., Miyao M., and Ikeguchi C. 2003]. As mentioned earlier, the word order in ASL is from general to specific and from large to small, similar to the Japanese language. This results in easier learning of each language.

We are currently working to expand our bilingual dictionary to a courseware dictionary. We are planning to test it and obtain feedback and suggestions. At first, we are going to expand it to conversational sentences, which will help users identify signs and their meanings within specific contexts.

6 Acknowledgements

We would like to thank Rie Nagoshi and Kyoko Tega for sharing their knowledge about JSL. We would also like to thank Tokyo Kasei Gakuin university for their financial support, and our seminar members for their assistance collecting data on Japanese and American sign language.

References

Suzuki, E., and Kakihana, K. Japanese and American Sign Language Dictionary System for Japanese and English Users, Proceedings of the LREC2002, pp.215-218,2002.

Hashimoto, T., 2000. The Report on the Present State of the Higher Education of the Hearing Impaired Persons in U.S.A.. – From the 8th Field Trip toGallaudent, RIT and NTID -(in Japanese). Tsukuba College of Technology Technical Report 2000.

Nakamura, K.(1). *"About American Sign Language,"* Web site : *http://www.deaflibrary.org/asl.html.*

Nakamura, K.(2). *"About Japanese Sign Language,"* Web site : *http://www.deaflibrary.org/jsl.html.*

Baker-Shenk.,Cokely C.and D., 1991. American Sign Language: Student text units 10-18. Washington, D.C.: Gallaudet University Press..

Yoden, Y. and Yamanoi K., 2000.The Groupware "Study Note" as a Tool for Digital Portfolio Assessment. *Proc. of 24th annual meeting of Japan Society for Science Education.*

Yonaiyama, A., and Ogata E., 2000. Easy Japanese Sign Language. *Natsume-sha.*

Pauly, M., Miyao M., and Ikeguchi C. 2003. Sign Language in the Language-teaching Classroom, The Proc. of JALT2003.

Building parallel corpora for eContent professionals

M. Gavrilidou, P. Labropoulou, E. Desipri, V. Giouli, V. Antonopoulos, S. Piperidis
Institute for Language and Speech Processing
Epidavrou & Artemidos 6
151 25 Maroussi, Greece.
{maria, penny, elina, voula, vantonop, spip} @ilsp.gr

Abstract

This paper reports on completed work carried out in the framework of the INTERA project, and specifically, on the production of multilingual resources (LRs) for eContent purposes. The paper presents the methodology adopted for the development of the corpus (acquisition and processing of the textual data), discusses the divergence of the initial assumptions from the actual situation met during this procedure, and concludes with a summarization of the problems attested which undermine the viability of multilingual parallel corpora construction.

1 Introduction

INTERA (Integrated European language data Repository Area, Contract 22076Y2C2DMAL2) is an EU-funded project within the eContent framework, aiming at

- building an integrated European Language Resources (LRs) area by connecting existing data centers at regional, national and international level, and
- at proposing "ways and techniques for LRs packaging to make it a profitable and attractive task to eContent professionals"; as an application of this task, the production of multilingual resources, namely parallel corpora and multilingual terminologies extracted from these, is undertaken (INTERA Technical Annex).

This paper focuses on the second aim of the project, presenting the work carried out in the area of parallel corpus production, identifying the steps followed in this process, in order to point out the problematic areas involved in the task and suggest ways of encompassing them.

2 Methodology and specifications

The process usually followed in the LRs production involves the following tasks: (a) identification of user needs and requirements, (b) specifications for the selection, construction and packaging of the LRs, (c) identification of potential sources, (d) construction of the LRs per se, (e) promotion and distribution of the LRs.

Given that INTERA is an eContent project, the target user group defined by the Technical Annex of the project was *eContent professionals and users*; furthermore, it was decided that the LRs to be produced (which would be of interest to this group) would be *parallel corpora* and *multilingual terminological lists*. Finally, the most important objective of the LRs production was the definition of a business model which would be attractive to the abovementioned target group.

The following sections discuss the actual steps taken for the implementation of these requirements.

The target group of eContent players addressed by the project has been further defined as consisting of professionals involved with the:

- production of digital content (authors or publishers)
- Globalization, Internationalization, Localization and Translation (GILT) processes, and
- development of Human Language Technology (HLT) software, ranging from multilingual information retrieval and extraction tools, to content management and Computer-Assisted Translation or Machine Translation solutions.

The next step concerned the identification of user needs and requirements on the basis of the professionals' working habits and processes. This was achieved by exploiting the results of a number of previous initiatives to roadmap the state-of-the-art in multilingual LRs, in combination with new initiatives undertaken in the framework of the project and targeted to the eContent world.

The surveys conducted in the framework of the ENABLER project (Maegaard et al. 2003, Gavrilidou & Desipri 2003) provided insights as to the existence and availability of different types of LRs, language demand, domains of interest, standards, etc. Although ENABLER focused on the LRs developer's point of view, a number of valuable results were elicited. Other surveys, such as those conducted by ELRA and its distribution

agency ELDA aiming at determining the needs of users with respect to available and potentially available LRs (http://www.elra.info/), or surveys available over the Internet through the sites of international organizations such as LISA and IDC or consultancy firms (http://www.globalsight.com, LISA 2001, LISA/AIIM 2001, LISA/OSCAR 2003) shed a light as to the availability of resources and relevant tools.

The information elicited from these surveys was coupled by a study of the activities of the eContent professionals as regards LRs, conducted in the framework of INTERA (Gavrilidou et al, 2004) through the circulation of a questionnaire distributed to potential users, as well as through personal contacts with a number of actors in the relevant fields. The main areas of the study concerned the types of LRs the eContent professionals are interested in, domains and languages of interest, and, most important, policies concerning the way they acquire, use and exploit LRs and tools.

The study of the target group yielded the following specifications:

- *domains*: it is obvious that eContent users are more interested in specialized domains than in general language resources; moreover, the survey results showed health/medicine, tourism, education, law, automotive industry and IT/telecommunications, as being the prevailing ones. In the framework of the INTERA project, however, we decided to focus on the prevailing domains as long as they promote multilingual and multicultural content. The selected domains are: *health, tourism, education and law*, which correspond to the predominant digital activities, namely, eTourism, eHealth, eLearning, eGovernment and eCommerce.
- *languages*: the focus of eContent and the needs of the users pointed towards the less widely spoken languages, including Balkan and Central and Eastern European languages (i.e the languages of the new EU countries).

  The project aims at the construction of a multilingual parallel corpus of 12 million words in total. The ideal scenario for the intended application of term extraction would be that of having a corpus with a source or pivot language and translations of the same texts in a number of target languages; however, given that the project aims at proposing realistic solutions to be adopted in the future by prospective LRs creators, real-life drawbacks should be taken into account; therefore, the limitations in the availability of existing resources (see section 3.1) dictated the decision to collect resources for four *pairs of languages*: Greek-English, Bulgarian-English, Slovene-English and Serbian-English.

The specifications for the processing of the corpus have been based on the requirements of its intended application, which is the *extraction of terminology*, and involve the following tasks:

- *alignment* of the texts: for the specific application purposes, alignment at sentence level has been deemed sufficient; however, the quality of the output is considered crucial; therefore, automatic processing is followed by human validation by language experts;
- external and internal *structural annotation*: the minimal requirements include segmentation at sentence level for the alignment task and metadata information that will be required for the distribution and re-use of the corpus;
- *linguistic processing*: below-Part of Speech (PoS) tagging and lemmatization is the minimum information required for the automatic term extraction task.

To ensure re-usability of the collected and processed material, compliance with the following internationally accredited standards was decided:

- the aligned material conforms to the TMX standard (Translation Memory eXchange, http://www.lisa.org/tmx/), which is XML-compliant. Being a vendor-neutral, open standard for storing and exchanging translation memories created by Computer Aided Translation (CAT) and localization tools, TMX standard was identified as a requirement for the eContent professionals. It allows easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process;
- for the external annotation, the IMDI metadata schema (IMDI, Metadata Elements for Session Descriptions, Version 3.0.4, Sept. 2003, http://www.mpi.nl/world/ISLE/schemas/schemas_frame.html) has been selected; the internal structural annotation adheres to the XCES standard, i.e. the XML version of the Corpus Encoding Standard (XCES, http://www.cs.vassar.edu/XCES/ and CES, http://www.cs.vassar.edu/CES/CES1-0.html).
- the linguistic annotation of the texts also adheres to the XCES standard, which incorporates the EAGLES guidelines for morphosyntactic annotation (http://www.ilc.cnr.it/EAGLES96/home.html).

3 Corpus construction

3.1 Text collection

In order to construct the parallel corpus, the first step consisted in the identification of potential sources, i.e. existing parallel corpora and, alternatively or additionally, textual material that could be used for the creation from scratch of the INTERA corpus.

Previous surveys (see section 2) that identify existing LRs as well as a search over the Internet attested the scarcity of available resources in the selected languages and domains, and so, the idea of re-using existing corpora was abandoned in favour of the construction of a new corpus from scratch.

The identification process of potential sources had to take into consideration the following requirements:

- to obtain texts from a variety of sources of interest to the eContent society,
- to ensure that the material was free of Intellectual Property Rights problems, either through the arrangement of specific agreements or by obtaining them from public sources.

The ideal candidates, in this respect, mainly consist of texts available over the Internet, provided by organizations/institutions that wish to make their own material available in more than one language, such as international organizations (e.g. United Nations, European Union, World Health Organization, Non-Governmental Organizations, etc.), multinational companies, companies with activities outside their own country (e.g. data describing company profiles & activities, product catalogues, etc.), public administration services (e.g. regarding bilateral agreements, regulations for immigrants, etc.), news agencies (targeting international broadcasting or for foreign language audience within their own country), official national government sites, national tourism organizations, etc. In all the above cases, the material consists of either web content per se (i.e. mainly bilingual web sites, rarely trilingual or quadrilingual) or of texts (official documents, technical reports, etc.) included in the web sites.

A more careful investigation, however, of web texts showed that although Internet is rapidly becoming multilingual, it is not yet parallel, especially as regards the languages involved in the project: most international bodies include original and translated texts but only in the more widely spoken languages. Moreover, a closer inspection of web texts that "seem" parallel, on the basis of structural similarities (e.g. similar size, paragraph segmentation, possible "anchors", such as list enumerators, etc.) showed that only sporadic parts of them were parallel. More problems arise from the fact that texts may contain large parts of foreign language material (e.g. EU regulations that include amendments to previous regulations by including the replacement text of specific paragraphs in all EU languages).

Given the above observations, cooperation with other data centers, with proven expertise in the area of LRs production for the specific project languages was sought; this would ensure content quality of the corpus, both during the selection (i.e. native speakers are better qualified to recognize true parallel material) and the encoding and validation processes, especially as regards the alignment validation and the linguistic processing. ILSP remains responsible for the construction of the Greek-English corpus, the collection and harmonization of the four subcorpora, the linguistic processing of the English texts and the addition of the IMDI metadata.

3.2 Text processing

Depending on the source that provided the original material (e.g. web site content, publishing house, translation company, etc.), different processing was required in order to arrive at the desired format adhering to the specifications set by the INTERA project; such as, indicatively:

- conversion of the original PDF/RTF/HTML etc. files into the format required by the various tools (tokenizer, aligner, tagger),
- cleanup of the texts from unwanted material (e.g. tables, figures, foreign language material, etc.)
- re-structuring of the original monolingual texts from the TMX file, when the source was the output of a Translation Memory,
- manual or semi-automatic annotation of metadata.

Each language team undertook the processing of the collected material (i.e. alignment and human validation, structural and linguistic annotation without human validation), using their own tools, thus ensuring that no time is lost over training with new tools and that the required language-dependent tools (especially taggers) used in the project are the most appropriate ones. The material to be delivered, however, at the end of all processes must be conformant to the selected standards.

The intervention of ILSP takes place only at the end of this process, with the purpose of validating the conformance of the results and of harmonizing any problematic issues. The most important point of this process is the linguistic annotation and, specifically, the harmonization of the different

tagsets used. In conformance with the methodology adopted in the project, i.e. of re-using existing material, whenever possible, with the least possible interventions, so as to ensure time and cost efficiency, it was decided to re-use only existing tools for each language, without making any modifications to the tools themselves but only conversion(s) of their output. Therefore, the task of harmonizing the output with regard to the morphosyntactic tags employed by each tagger is the last stage of the procedure, where all tagsets are mapped to one, based on the EAGLES guidelines.

4 Conclusions

In this paper, we described the methodology followed in the construction of a multilingual parallel corpus; this task has been interpreted as a test application endeavor in the process of defining a business model for the LRs production. The effort was to identify gaps and shortcomings in the process usually employed by LRs producers (or users who might wish to create their own LRs) and to suggest ways of remedying them. Our findings include:

- *problems faced during the acquisition phase*: although an increasing supply of raw data (e.g. over Internet) and tools capable of exploiting this data (e.g. web crawlers that can identify and download texts in a given language) is attested, there is also a need for the enhancement of these tools with more intelligent techniques (e.g. incorporation of alignment techniques during the acquisition process in order to spot potential parallel texts, identification and mark-up of large foreign language excerpts),
- *problems faced during the processing phase*: in order to enhance the LRs production effort, the re-use of existing tools is considered crucial. It is true that an increasing number of tools are available for text processing; however, this is oriented mainly towards the major languages. Moreover, information concerning the existence, availability and operation of existing tools is not easy to locate – a gap that the other pillar of INTERA tries to remedy through the building of an integrated European Language Resources area. Additionally, tools must be enhanced with respect to two directions: improvement of the tools themselves (e.g. more robust alignment techniques) and interoperability of all relevant tools currently used at different phases of processing. The issue of interoperability is closely related with the issue of standards. The promotion and deployment of existing standards as well as the creation of new standards, when these are lacking, is important to ensure viability and re-use of LRs, given the cost of their production.

References

Gavrilidou, M., E. Desipri. 2003. Final Version of the Survey, ENABLER Deliverable 2.1.

Gavrilidou, M. E. Desipri, P. Labropoulou, S. Piperidis, N. Calzolari, M. Monachini & C. Soria. 2004. Technical specifications for the selection and encoding of multilingual resources, INTERA (Integrated European language data Repository Area), Deliverable 5.1.

IMDI, Metadata Elements for Session Descriptions, Version 3.0.4, Sept. 2003.

INTERA – eContent 2002 Integrated European languages data Repository Area, Technical Annex.

LISA. 2001. The LISA Globalization Strategies Awareness Survey.

LISA/AIIM. 2001. The Black Hole in the Internet: LISA/AIIM Globalization Survey.

LISA/OSCAR. 2003. Translation Memory Survey.

Maegaard, B., K. Choukri, V. Mapelli, M. Nikkhou & C. Povlsen. 2003. Language resources-Industrial needs, ENABLER Deliverable 4.2.

Revising the WORDNET DOMAINS Hierarchy: semantics, coverage and balancing

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, Emanuele Pianta
ITC-irst – Istituto per la Ricerca Scientifica e Tecnologica
Via Sommarive 18, Povo – Trento, Italy, 38050
email:{bentivo, forner, magnini, pianta}@itc.it

Abstract

The continuous expansion of the multilingual information society has led in recent years to a pressing demand for multilingual linguistic resources suitable to be used for different applications.

In this paper we present the WordNet Domains Hierarchy (WDH), a language-independent resource composed of 164, hierarchically organized, domain labels (e.g. Architecture, Sport, Medicine). Although WDH has been successfully applied to various Natural Language Processing tasks, the first available version presented some problems, mostly related to the lack of a clear semantics of the domain labels. Other correlated issues were the coverage and the balancing of the domains. We illustrate a new version of WDH addressing these problems by an explicit and systematic reference to the Dewey Decimal Classification. The new version of WDH has a better defined semantics and is applicable to a wider range of tasks.

1 Introduction

The continuous expansion of the multilingual information society with a growing number of new languages present on the Web has led in recent years to a pressing demand for multilingual applications. To support such applications, multilingual language resources are needed, which however require a lot of human effort to be built. For this reason, the development of language-independent resources which factorize what is common to many languages, and are possibly linked to the language-specific resources, could bring great advantages to the development of the multilingual information society.

A language-independent resource, usable in many automatic and human applications, is represented by *domain hierarchies.* The notion of domain is related to similar notions such as *semantic field, subject matter, broad topic, subject code, subject domain, category.* These notions are used, sometimes interchangeably, sometimes with significant distinctions, in various fields such as linguistics, lexicography, cataloguing, text categorization. As far as this work is concerned, we define a *domain* as an area of knowledge which is somehow recognized as unitary. A domain can be characterized by the name of a discipline where a certain knowledge area is developed (e.g. chemistry) or by the specific object of the knowledge area (e.g. food). Although objects of knowledge and disciplines that study them are clearly related, the relation between these two points of view on domains is sometimes blurred and may be a source of uncertainty on their exact definition.

Another interesting duality when speaking about domains is related to the fact that knowledge manifests itself in both words and texts. So the notion of domain can be applied both to the study of words, where a domain is the area of knowledge to which a certain lexical concept belongs, or to the study of texts, where the domain of a text is its broad topic. In this work we will assume that also these two points of view on domains are strictly intertwined.

By their nature, domains can be organized in hierarchies based on a relation of specificity. For instance we can say that TENNIS is a more specific domain than SPORT, or that ARCHITECTURE is more general than TOWN PLANNING.

Domain hierarchies can be usefully integrated into other linguistic resources and are also profitably used in many Natural Language Processing (NLP) tasks such as Word Sense Disambiguation (Magnini et al. 2002), Text Categorization (Schutze, 1998), Information Retrieval (Walker and Amsler, 1986).

As regards the usage of Domain hierarchies in the field of multilingual lexicography, an example is given by the EuroWordNet Domain-ontology, a language independent domain hierarchy to which interlingual concepts (ILI-records) can be assigned (Vossen, 1998). In the same line, see also the SIMPLE domain hierarchy (SIMPLE, 2000).

Large domain hierarchies are also available on the Internet, mainly meant for classifying web documents. See for instance the Google and Yahoo directories.

A large-scale application of a domain hierarchy to a lexicon is represented by WORDNET DOMAINS (Magnini and Cavaglià, 2000). WORDNET DOMAINS is a lexical resource developed at ITC-irst where each WordNet synset (Fellbaum, 1998) is annotated with one or more domain labels

selected from a domain hierarchy which was specifically created to this purpose. As the WORDNET DOMAINS Hierarchy (WDH) is language-independent, it has been possible to exploit it in the framework of MultiWordNet (Pianta et al., 2002), a multilingual lexical database developed at ITC-irst in which the Italian component is strictly aligned with the English WordNet. In MultiWordNet, the domain information has been automatically transferred from English to Italian, resulting in a Italian version of WORDNET DOMAINS. For instance, as the English synset {court, tribunal, judicature} was annotated with the domain LAW, also the Italian synset {corte, tribunale}, which is aligned with the corresponding English synset, results automatically annotated with the LAW domain. This procedure can be applied to any other WordNet (or part of it) aligned with Princeton WordNet (see for instance the Spanish WordNet).

It is worth noticing that two of the main on-going projects addressing the construction of multilingual resources, that is MEANING (Rigau et al. 2002) and BALKANET (see web site), make use of WORDNET DOMAINS. Finally, WORDNET DOMAINS is being profitably used by the NLP community mainly for Word Sense Disambiguation tasks in various languages.

Another application of domain hierarchies can be found in the field of *corpus creation*. In many existing corpora (see for instance the BNC, the ANC, the Brown and LOB Corpora) domain is one of the most used criteria for text selection and/or classification. Given that a domain hierarchy is language independent, if the same domain hierarchy is used to build reference corpora for different languages, then it would be easy to create (a first approximation of) *comparable corpora* by putting in correspondence corpora sections belonging to the same domain.

An example of a corpus in which the complete representation of domains is pursued in a systematic way is represented by the MEANING Italian corpus, a large size corpus of written contemporary Italian in which a subset of the WDH labels has been chosen as the fundamental criterion for the selection of the texts to be included in the corpus (Bentivogli et al., 2003).

Given the relevance of language-independent domain hierarchies for multilingual applications, it is of primary importance that these resources have a well-defined semantics and structure in order to be useful in various application fields. This paper reports the work done to improve the WDH so that it complies with such requirements. In particular, the WDH revision has been carried out with reference to the Dewey Decimal Classification.

The paper is organized as follows. Section 2 briefly introduces the WORDNET DOMAINS Hierarchy and its main characteristics, with a short overview of the Dewey Decimal Classification system. Section 3 describes features and properties of the revision. Finally, in section 4, conclusions are reported.

2 The WordNet Domains Hierarchy

The first version of the WDH was composed of 164 domain labels selected starting from the subject field codes used in current dictionaries, and the subject codes contained in the Dewey Decimal Classification (DDC), a general knowledge organization tool which is the most widely used taxonomy for library organization purposes.

Domain labels were organized in five main trees, reaching a maximum depth of four. Figure 1 shows a fragment of one of the five main trees in the WORDNET DOMAINS original hierarchy.

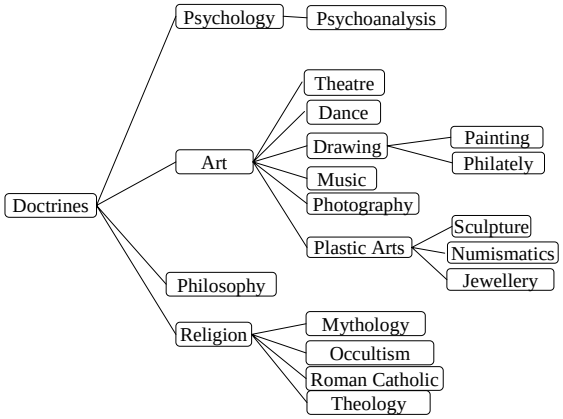


Figure 1: Fragment of the original WDH

Domain labels were initially conceived to be application-oriented, that is, they have been integrated in WordNet with the main purpose of allowing the categorization of word senses and to provide useful information during the disambiguation process.

The second level of WDH, where the so-called *Basic Domains* are represented, includes labels such as ART, SPORT, RELIGION and HISTORY, while in the third level a degree of major specialization is reproduced, and domains, like for example, DRAWING, PAINTING, TENNIS, VOLLEYBALL, and ARCHAEOLOGY can be found. For NLP tasks, the set of *Basic Domains* has proved to possess a suitable level of abstraction and granularity.

Although the first version of WDH found many applications in different scenarios, it presented some problems. First, the domain labels did not have a defined semantics. The content of the labels

could be suggested by the lexical meaning of their name, but there was no explicit indication about their intended interpretation.

Second, it was not clear whether the *Basic Domains* met certain requirements such as knowledge coverage and balancing. In fact, the *Basic Domains* are supposed to possess a comparable degree of granularity and, at the same time, to cover all human knowledge. However, they did not always posses such characteristics. For instance VETERINARY was put at the same level as ECONOMY, although these two domains obviously do not posses the same level of granularity. Moreover not all branches of human knowledge were represented (see for instance the HOME domain).

The purpose of the work presented here was, therefore, to find a solution for such problems, in order to improve the applicability of WDH in a wider range of fields. The solution we propose is crucially based on the Dewey Decimal Classification (edition 21), which has been used as a reference point for defining a clear semantics, preventing overlapping among domains, and assessing the *Basic Domains* coverage and granularity issues.

2.1 The Dewey Decimal Classification (DDC)

The Dewey Decimal Classification (DDC) system (Mitchell et al. 1996) is the most widely used taxonomy for library classification purposes providing a logical system for the organization of every item of knowledge through well-defined subject codes hierarchically organized. The semantics of each subject code is determined by a numeric code, a short lexical description associated to it, and by the hierarchical relations with the other subject codes. Another characteristic of the DDC is that a handbook is available explaining how texts should be classified under subject codes.

The DDC is not just for organizing book collections; it has also been licensed for cataloguing internet resources (see for example BUBL http://bubl.ac.uk/link/) and it was conceived to accommodate the expansion and evolution of the body of human knowledge.

The DDC hierarchy is arranged by disciplines (or fields of study), and this entails that a subject may appear in more than one discipline, depending on the aspect of the topic discussed.

The DDC hierarchical structure allows a topic to be defined as part of the broader topic above it, and that determines the meaning of the class and its relation to other classes. At the broadest level, called *Main Classes* (or *First summary)*, the DDC is composed of ten mutually exclusive main classes, which together cover the entire world of knowledge. Each main class is sub-divided into ten *divisions*, (the *Hundred Divisions*, or *Second Summary*) and each division is split into ten *sections* (the *Thousand Section*, also called *Third Summary*).

Each category in the DDC is represented by a numeric code as the example below shows.

700 *Art*
7**3**0 *Plastic Arts*
*73**6** Carving*
*736.**2** Precious Stones*
*736.2**3** Diamonds*
*736.2**5** Sapphires*
*736.**4** Wood*
*73**8** Ceramic Arts*
*73**9** Art Metalwork*
7**4**0 *Drawing*
7**5**0 *Painting*

The first digit of the numbers indicates the main class, (**7**00 is used for all *Arts*) the second digit indicates the hundred division, (7**3**0 corresponds to *Plastic arts*, 7**4**0 to *Drawing*, 7**5**0 to *Painting*) and the third digit indicates the section (73**6** represents *Carving*, 73**8** *Ceramic arts*, 73**9** *Art metalwork*). Moreover, almost all sub-classes are further subdivided. A decimal point follows the third digit until the degree of specification needed (736.2**3** *Diamonds*, 736.2**5** *Sapphires*).

3 The Revision of the WDH

The revision of the first version of the WDH aimed at satisfying the following properties and characteristics:

- *semantics*: each WDH label should have an explicit semantics and should be unambiguously identified;
- *disjunction*: the interpretation of all WDH labels should not overlap;
- *basic coverage*: all human knowledge should be covered by the *Basic Domains*;
- *basic balancing*: most *Basic Domains* should have a comparable degree of granularity.

In the following sections we are going to show how a systematic mapping between WDH and DDC can be used to enforce each of the above characteristics.

3.1 Semantics

To give the domain labels a clear semantics so that they can be unambiguously identified and interpreted, we decided to associate each domain label to one or more DDC codes as shown below in Table 1.

WDH Domains	DDC Codes
Art	[700-(790-(791.43,792,793.3), 710,720,745.5)]
Plastic arts	730
Sculpture	[731:735]
Numismatics	737
Jewellery	739.27
Drawing	[740-745.5]
Painting	750
Graphic arts	760
Philately	769.56
Photography	770
Music	780
Cinema	791.43
Theatre	[792-792.8]
Dance	[792.8,793.3]

Table 1: Fragment of the new WDH with the respective DDC codes

In many cases we found a one-to-one mapping between a WDH label and a DDC code (e.g. PAINTING mapped onto 750 or CINEMA onto 791.43). When one-to-one mappings were not found, artificial DDC codes were created. An artificial code, represented within square brackets, is created with reference to various DDC codes or parts of them. To describe artificial nodes, certain conventions have been adopted.

(i) A series of non-consecutive codes is listed separated by a comma (see DANCE).

(ii) A series of consecutive codes is indicated by a range. For instance, the series [731, 732, 733, 734, 735] is abbreviated as [731:735] (see SCULPTURE).

(iii) A part of a tree is represented as the difference between a tree and one or more of its subtrees, where the tree and the subtrees are identified by their roots (see DRAWING).

(iv) The square brackets should be interpreted as meaning "the generalities" of the composition of codes contained in the brackets. So, for instance, [731:735] should be interpreted as the generalities of the codes going from 731 to 735. In the original DDC, generalities are identified by the 0 decimal. For instance, the code 700 refers to the generalities of the codes from 710 to 790.

To establish a mapping between labels and codes we exploited the names of the DDC categories and their description in the DDC manual. This worked pretty well in most cases, but there are some exceptions. Take for instance the TOURISM domain. Apparently tourism does not occur as a category in the DDC. On a closer inspection it came out that the categories which are most clearly related to tourism are 910.202:*World travel guides* and 910.4:*Accounts of travel*.

Note that a WDH domain can be mapped onto codes included in different DDC main classes, i.e. disciplines. For example ARTISANSHIP (745.5:*Handicrafts*, 338.642:*Small business*) maps onto categories located partly under 700:*Art* and partly under 300:*Social Sciences*. The same happens with SEXUALITY, a domain that following the DDC is studied by many different disciplines, e.g. philosophy, medicine, psychology, body care.

As a consequence of the systematic specification of the semantics of the WDH domains, some of them have been re-labeled with regard to the previous version of the hierarchy. For instance, the domain BOTANY has been changed to PLANTS, ZOOLOGY to ANIMALS, and ALIMENTATION to FOOD. This change of focus from the name of the discipline to the name of the object of the discipline is not only in compliance with the new edition of the DDC, but it also reflects current and international usage (see, for example, Google categories). In some cases the change of the domain name comes along with a change of its intended interpretation. For instance, we have decided to enlarge the semantics of the domain ZOOTECHNICS and to call it ANIMAL HUSBANDRY, a more generic domain which was missing in the previous hierarchy.

In most cases the hierarchical relations between the WDH domains are the same as the relations holding between the corresponding DDC codes: MUSIC is more specific than ART in the same way as 780:*Music* is more specific than 700:*The Arts*. To reinforce the hierarchical parallelism between the WDH and the DCC, we re-located some domains with regard to the previous WDH hierarchy. For example, OCCULTISM, which was placed under RELIGION in the old hierarchy, has been moved under the newly created domain PARANORMAL. Also, TOPOGRAPHY, previously placed under ASTRONOMY, has now been moved under GEOGRAPHY.

In a few cases however we did not respect the hierarchical relations specified by the DDC, as in the case of the ARCHITECTURE domain shown in Table 2. ARCHITECTURE has been mapped onto 720:*Architecture* and TOWN PLANNING onto 710:*Civic & landscape art*.

WDH Domains	DDC Codes
Architecture	[645,690,710,720]
Town Planning	710
Buildings	690
Furniture	645

Table 2: A fragment of WDH for ARCHITECTURE

However, whereas the 710 code is sibling of 720 in the DDC, TOWN PLANNING is child of ARCHITECTURE in WDH. Also, ARCHITECTURE and TOWN PLANNING should be under ART according to the DDC, but they have been placed under APPLIED SCIENCE in WDH.

3.2 Disjunction

This property requires that no DDC code is associated to more than one WDH label. In only one case this requirement has not been met. Apparently, the DDC does not distinguish between the disciplines of Sociology and Anthropology, and reserves the codes that go from 301 to 307 to both of them. Although these two disciplines are strictly connected, it seems to us that in the current practice they are considered as distinct. So the WDH contains two distinct domains for SOCIOLOGY and ANTHROPOLOGY, which partially overlap because they both map onto the same DDC codes 301:307.

3.3 Basic Coverage

The term *basic coverage* refers to the ideal requirement that all human knowledge be covered by the totality of the *Basic Domains* (i.e. the domains composing the second level of WDH). Also in this case, we used the DDC as a gold standard to measure the coverage of WDH. Given the fact that the DDC has been used for more than a century to classify books and written documents all over the world, we can assume that the DDC guarantees a complete representation of all branches of knowledge. So the *basic coverage* has been manually checked by verifying that all (or almost all) the DDC categories can be assigned to at least one *Basic Domain*.

From a practical point of view, it would be very complicated to check all the thousands of codes contained in the DDC. Thus, our check relied on two assumptions. First, when the *Basic Domains* are taken as a stand alone set, the semantics of a *Basic Domain* is given by its specific code together with the codes of its subdomains. Second, once a DDC code is covered by a *Basic Domain*, inductively, all the more specific categories are covered as well. These assumptions allowed us to actually check only the topmost DDC codes. For example, let's take the **3**00 main class of the DDC. Table 3 below shows that all the sub-codes of the **3**00 class are covered by one or more domains.

In order to improve the overall WDH coverage, 5 completely new domains have been introduced (the first three are *Basic*): PARANORMAL, HOME, HEALTH, FINANCE and GRAPHIC ARTS.

Codes	DDC Categories	WDH Domains
300	• *Social sciences*	• SOCIAL SCIENCE • SOCIOLOGY • ANTHROPOLOGY
310	• *General statistics*	• SOCIOLOGY
320	• *Political science*	• POLITICS
330	• *Economics*	• ECONOMY
340	• *Law*	• LAW
350	• *Public administration & military service*	• ADMINISTRATION • MILITARY
360	• *Social problems & services*	• SOCIOLOGY • ECONOMY • SEXUALITY
370	• *Education*	• PEDAGOGY
380	• *Commerce, communication, transport*	• COMMERCE • TELECOMMUNICATION • TRANSPORT
390	• *Customs, etiquette, folklore*	• FASHION • ANTHROPOLOGY • SEXUALITY

Table 3: Coverage of the **3**00 DDC class

We can now assume that the domain-coverage of the new version of WDH is almost equivalent to that of the DDC, thus ensuring the complete representation of all branches of knowledge.

The new WDH allowed us to fix a number of synset classifications that were unsatisfactory in the previous version of WORDNET DOMAINS. For instance, in the first version of WORDNET DOMAINS the English/Italian synset {microwave oven, microwave}/{forno a microonde, microonde} was annotated with the FURNITURE domain, while the synset {detergent}/{detersivo} was annotated with FACTOTUM (i.e. no specific domain) as no better solution was available. The new WDH hierarchy allows for a more appropriate classification of both synsets within the new HOME domain.

A few DDC codes are not covered by the new list of domains either. These are the codes under the 000:*Generalities* class which includes disciplines such as 010:*Bibliography*, 020:*Library & information sciences*, 030:*Encyclopedic works*, 080:*General collections*. This section has been specifically created for cataloguing general and encyclopedic works and collections. So it is a idiosyncratic category which is not based on subject but on the genre of texts.

Another set of codes which remains not covered by WDH are those going from 420 to 490 and from 810 to 890. These DDC codes are devoted to specific languages and literatures of different countries, for example, 430:*Germanic Languages*, 440:*Romance Languages*, 810:*American Literature in English*, etc. These codes are undoubtedly relevant for the classification of books, but are not compatible with the rationale of WDH, which is meant to be a language-independent resource.

3.4 Basic Balancing

The requirement about *basic balancing* is meant to assure that all *Basic Domains* have a comparable degree of granularity.

Defining a granularity metrics for domains is a complex issue, for which only a tentative solution is provided here. At a first glance, three aspects could be taken into consideration: the number of publications about a domain, the number of sub-codes in the DDC, and the relevance of a domain in the social life.

As a first attempt, balancing could be evaluated referring to the number of publications classified under each *Basic Domain*. In fact, data are available about the number of texts classified under each of the DDC codes. Unfortunately, the number of books published under a certain category may not be indicative of its social relevance: very specialized domains may include a high number of publications, which however circulate in a restricted circle, with low social impact. For example, the number of texts classified in the History domain turns out to be more then ten times the number of texts catalogued under the Computer Science domain. However, if one looks at the number of HTML pages available on the Internet, or the number of magazines sold in a newspaper stand, or the number of terms used in everyday life, one cannot maintain that History is ten times more relevant than Computer Science.

Another approach for evaluating the granularity of domains could be to take into account the number of DDC sub-codes corresponding to each *Basic Domain*. Unfortunately, also this approach gives results which are far from being satisfactory. The fact that a discipline has many subdivisions seems not to be clearly correlated with its relevance. For instance in the DDC manual (version 21) 105 pages can be put in correspondence with the ENGINEERING domain, whereas only 26 correspond to SPORT. It should also be said that there is no correlation between the number of publications and the number of sub-categories in the DDC. For instance, ARCHITECTURE has a great number of publications classified under it, but on the contrary, the number of sub-categories in the DDC is very limited.

The third criterion to evaluate the granularity of domains is their social relevance, which seems not to be captured adequately by the previous two criteria. Of course, social relevance is very difficult to evaluate. We tentatively took into consideration the organization of Internet hierarchies such as the Google and Yahoo directories, which seem to be closer than the DDC to represent the current social relevance of certain domains. See for instance the huge number of HTML pages classified in Google under the topic *Television Programs*. Of course Internet is only a partial view of the organization of human knowledge, so we cannot simply rely on the Internet to evaluate the granularity of the domains.

None of the approaches analyzed so far seems to fit our needs. Thus we took into consideration a fourth criterion, which is based on the DDC as well. Instead of counting the number of subdivisions under a certain DDC code, we measured the depth of the code from the top of the hierarchy. For instance we can say that 700:*Art* has depth 1, 780:*Music* has depth 2, 782:*Vocal Music* has depth 3, and so on. We make the assumption that two DDC codes with the same depth have the same granularity. For instance we assume that 782:*Vocal Music* and 382:*Foreign Trade* have the same granularity (both have depth 3).

In order to evaluate the granularity of the *Basic Domains* against the DDC, we can compare WDH labels and DDC codes with the same depth. Given that the *Basic Domains* have depth 2, we should compare them to the so called *Hundred Divisions* (000, 010, 020, 030, …, 100, 110, 120, etc.). Summing up, we will say that the *Basic Domains* are balanced if they can all be mapped onto the *Hundred Divisions*. Also, in the comparison we should take into account that the *Basic Domains* are 45, whereas the *Hundred Divisions* are 100. So, we expect that in the average, one *Basic Domain* maps onto two *Hundred Divisions* with a small degree of variance with respect to the average.

What we have obtained from the analysis of the new WDH is the following: out of 45 *Basic Domains*

- 4 domains map onto a *Main Class* (depth 1)
- 18 domains are mapped at the *Hundred Divisions* level (depth 2)
- 6 domains are mapped at different DDC levels, with the majority of DDC codes at depth 2
- 17 domains map onto subdivisions of depth 3 and 4.

As for the average number of DDC codes covered by each *Basic Domain*, the variance is quite high. Certain *Basic Domains* cover a big number of codes from the *Hundred Divisions*. For instance HISTORY, and ART cover 6 codes each. Instead, in most cases, one *Basic Domain* covers only one DDC code (e.g. LAW and 340:*Law*).

The evaluation of the granularity of the *Basic Domains* according to the proposed criterion can be considered satisfactory even if the results diverge somewhat from what expected in principle.

To explain this partial divergence in the granularity of domains, one should take into

consideration that the DDC has been created relying heavily on the academic organization of knowledge disciplines. On the other side, in the practical WDH reorganization process we tried to balance somehow this discipline-oriented approach, by taking into account also the social relevance of domains. This has been done by relying on the organization of Internet directories and on our personal intuitions.

Such an approach led us to put at the *Basic* level WDH labels corresponding to DDC codes with depth higher than 2 (more specific than the *Hundreds Divisions*). See for instance the positioning of RADIO+TV, FOOD, HEALTH, and ENVIRONMENT at the *Basic* level, even if they correspond to DDC codes of level 3 and 4. Instead, ANIMALS and PLANTS were not *Basic* in the previous version of WDH, but have been promoted to the *Basic* level in accordance with the granularity level they have in the DDC.

Other domain labels have been placed at a lower level then expected with reference to the DDC. For instance PHILOSOPHY, ART, RELIGION, and LITERATURE have been put at the *Basic* Level, even if they correspond to DDC codes belonging to the *Main Classes* (depth 1). On the other side ASTROLOGY, ARCHAEOLOGY, BODY CARE, and VETERINARY which were *Basic* in the previous version of the WDH, have been demoted at a lower level in accordance with the granularity they have in the DDC. Only in one case this process of demotion has led to the elimination of a sub-domain, that is TEXTILE.

4 Conclusions

In this paper we described the revision of the WORDNET DOMAINS Hierarchy (WDH), with the aim of providing it with a clear semantics, and evaluating the coverage and balancing of a subset of the WDH, called *Basic Domains*. This has been done mostly by relying on the information available in the Dewy Decimal Classification (DDC). A semantics has been provided to the WDH labels by defining one or more pointers to DDC codes. The coverage of the *Basic Domains* has been evaluated by checking that each DDC code is covered by at least one *Basic Domain*. Finally, balancing has been evaluated mostly by comparing the granularity of the *Basic Domains* with the granularity of a subset of the DDC called the *Hundred Divisions*. Balancing is the aspect of the *Basic Domains* which diverges more clearly from the DDC. This is explained by the fact that we took in higher consideration the social relevance of domains.

We think that the new version of the WDH is better suited to act as a useful language-independent resource in the fields of computational lexicography, corpus building, and various NLP applications.

5 Acknowledgements

Thanks to Alfio Gliozzo for his useful comments and suggestions about how to improve the WORDNET DOMAINS Hierarchy.

References

BALKANET http://www.ceid.upatras.gr/Balkanet/

L. Bentivogli, C. Girardi and E. Pianta. 2003. The MEANING Italian Corpus. In *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, United Kingdom.

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press, Boston.

B. Magnini and G. Cavaglià. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC-2000*. Athens, Greece.

B. Magnini, C. Strapparava, G. Pezzulo and A. Gliozzo. 2002. The Role of Domain Information in Word Sense Disambiguation. *Journal of Natural Language Engineering (Special Issue on evaluating Word Sense Disambiguation Systems)*, 9(1):359:373.

J.S. Mitchell, J. Beall, W.E. Matthews and G.R. New (eds). 1996. *Dewey Decimal Classification* Edition 21 (DDC 21). Forest Press, Albany, New York.

E. Pianta, L. Bentivogli and C. Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First Global WordNet Conference*. Mysore, India.

G. Rigau, B. Magnini, E. Agirre, P. Vossen and J. Carrol. 2002. MEANING: a Roadmap to Knowledge Technologies. In *Proceedings of the COLING-2002 workshop "A Roadmap for Computational Linguistics"*. Taipei, Taiwan.

H. Schutze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97-123.

SIMPLE. 2000. *Linguistic Specifications*. Deliverable D2.1, March 2000.

P. Vossen (ed). 1998. *Computers and the Humanities (Special Issue on EuroWordNet)*, 32(2-3).

D.E. Walker and R.A. Amsler. 1986. *Analyzing Language in Restricted Domain. Sublanguage description and Processing*. Lawrence Earlbaum, Hillsdale NJ.

Appendix : The first two levels of the WDH new version with the corresponding DDC codes

TOP-LEVEL	BASIC DOMAINS	DDC
Humanities		
	History	[920:990]
	Linguistics	410
	Literature	[800, 400]
	Philosophy	[100-(130, 150, 176)]
	Psychology	150
	Art	[700-(710, 720, 745.5, 790-(791.43, 792, 793.3))]
	Paranormal	130
	Religion	200
Free_Time		[790-(791.43, 792, 793.3)]
	Radio-Tv	[791.44, 791.45]
	Play	[793.4:795-794.6]
	Sport	[794.6, 796:799]
Applied_Science		600
	Agriculture	[338.1, 630]
	Food	[613.2, 613.3, 641, 642]
	Home	[640-(641, 642, 645)]
	Architecture	[645, 690, 710, 720]
	Computer_Science	[004:006]
	Engineering	620
	Telecommunication	[383, 384]
	Medicine	[610-(611, 612, 613)]
Pure_Science		500
	Astronomy	520
	Biology	[570-577, 611, 612-612.6]
	Animals	590
	Plants	580
	Environment	577
	Chemistry	540
	Earth	[550, 560, 910-(910.4, 910.202)]
	Mathematics	510
	Physics	530
Social_Science		[300.1:300.9]
	Anthropology	[301:307, 395, 398]
	Health	[613-(613.2, 613.3, 613.8, 613.9)]
	Military	[355:359]
	Pedagogy	370
	Publishing	070
	Sociology	[301:319-(305.8, 306.7), 360-(363.4, 368)]
	Artisanship	[338.642, 745.5]
	Commerce	[381, 382]
	Industry	[338-(338.1, 338.642), 660, 670, 680]
	Transport	[385:389]
	Economy	[330-(334, 338), 368, 650]
	Administration	[351:354]
	Law	340
	Politics	320
	Tourism	[910.202, 910.4]
	Fashion	[390-(392.6, 395, 398), 687]
	Sexuality	[155.3, 176, 306.7, 363.4, 392.6, 612.6, 613.96]
	Factotum	

PolyphraZ : a tool for the management of parallel corpora

Najeh HAJLAOUI
GETA, CLIPS, IMAG
Université Joseph Fourier, BP 53
38041 Grenoble, France
Najeh.Hajlaoui@imag.fr

Christian BOITET
GETA, CLIPS, IMAG
Université Joseph Fourier, BP 53
38041 Grenoble, France
Christian.Boitet@imag.fr

Abstract

The PolyphraZ tool is being developed in the framework of the TraCorpEx project (Translation of Corpora of Examples), to manage parallel multilingual corpora through the web. Corpus files (monolingual or multilingual) are firstly converted to a standard coding (CXM.dtd, UTF8). Then, they are assembled (CPXM.dtd) to visualize them in parallel through the web. In a third stage, they are put in a Multilingual Polyphraz Memory (MPM). A "polyphrase" is a structure containing an original sentence and various proposals of equivalent sentences, in the same and other languages. An MPM stores one or more corpora of polyphrazes. The MPM part of PolyphraZ has 3 main web interfaces. One is a web-oriented translator workstation (TWS), where suggestions or translations come from the MPM itself, which functions as its own translation memory, and from calls to MT systems. Another serves to send sentences to MT systems with appropriate parameters, and to run various evaluation measures (NIST, BLEU, and distance computations) in order to propose to the translator a "best" proposal. A third interface is planned for giving feedbacks to the developers of the MT systems, in the form of lists of unknown or wrongly translated words, with suggestions for correct translations, and of parallel presentation of pairs of translations showing the "editing work" to be done to get one from the other. The first 2 stages are operational, and used for experimentation and MT evaluation on the CSTAR 5-lingual BTEC corpus and on the Japanese-English Tanaka corpus used as a source of examples in electronic dictionaries (JDict, Papillon). A main goal of this effort is to offer occasional and volunteer translators and posteditors access to a free TWS and to sharable translation memories put in the MPM format.

1 Introduction

Due to Internet grow, the number of available documents grows dramatically. There is a strategic need for companies to produce and manage information written in more than 30 languages (HP, IBM, MS, Caterpillar). This requires powerful tools to manage multilingual documents.

Current techniques for handling multilingual documents use large-grained linking (at the level of HTML pages), but don't allow fine-grained synchronization (at paragraph or sentence level) and don't permit bilingual or multilingual editing through the Web.

The interest to synchronize at least at the level of sentences is double:

- make it possible to use Machine Aided Human Translation (MAHT) techniques, in particular translation memories, for translating and postediting multilingual documents.
- add UNL tags at sentence level to store the translations as well as UNL hypergraphs (anglosemantic interlingual representations), from which raw (or rough!) translations into other languages can be obtained from distant "deconversion" servers.

Here, we are not concerned with the problem of aligning parallel monolingual documents, or realigning them after they have been modified, a frequent need in the case of leaflets and booklets. (Assimi,2000) proposed a tool to handle the non-centralized management of the evolution of multilingual parallel documents. We consider the case, frequent in the industry, where documents are managed centrally, even if they are distributed on several sites. What happens in general is that they are aligned at the level of large blocks, with one file per block and language (fileXXX.en.htm, fileXXX.fr.htm etc. for HTML pages).

What we propose is to align them at the level of sentences, but of course not to have one file per sentence. Rather, if there are N languages, for a given "block" corresponding to some unit of processing (e.g. visualization), we will have either

N monolingual sentence-aligned files, or 1 multilingual file. In both cases, sentences or place holders for sentences will be linked to a MPM to manage translation and postedition.

We began to build PolyphraZ in the context of the TraCorpEx project (Translation of Corpora of Examples). A more recent motivation is to extend the BTEC corpus of CSTAR III (163000 sentences in tourism) to French and Arabic, and to evaluate various Chinese-English MT systems on it.

We will first present the data we start with, and our goals in more detail. In a second part, we will describe the architecture of PolyphraZ, starting from scenarios of use and types of users. Lastly, we will describe the current status of this work.

2 TraCorpEx and PolyphraZ

2.1 Context

The TraCorpEx project has several contexts: the Papillon project (Papillon) of co-operative construction of a large multilingual lexical base on the Web, the C-STAR III project (C-STAR III) of translation of spoken dialogues, a French and Tunisian project (Hajlaoui, Boitet, 2003b), the UNL project (UNL) of communication and multilingual information system, and the PhD research of the various participants in this project.

2.2 Current data and problems

We have initially 2 "parallel" corpora, structured differently.

- The BTEC corpus of C-STAR is made of 5 sets of 163 files of 12K to 40K, each containing 1000 sentences, in English, Japanese (coded in EUC), Chinese and Korean, for a total of 6.1 Mo per language.
- The TANAKA corpus (Japanese-English), given to the Papillon project a few months before the death of its author in 2002, is made of 45 files for a total of 18.4 Mo. It contains sentences of newspapers or teaching works of NHK for the training of English by the Japanese. Each file is bilingual.

We have also corpora from the UNL project, where each document is a multilingual file containing for each sentence its text in source language, a UNL graph, the result of deconversions in a certain number of languages, and possibly their revisions, or direct manual translations.

All these "parallel" corpora are aligned at the level of sentences. As it would be interesting to show correspondences at finer levels (syntagms, chunks, words), we design PolyphraZ to later add tools for subsential alignement such as the one developed by Ch. Chenon for his Ph.D.

In other corpora, we may be obliged to go up to the level of paragraphs, because sentences will not be aligned perfectly. That will not be done completely in PolyphraZ, but at the level of the structure of the multilingual document itself: if 2 sentences are translated by 3, each of the 5 sentences will be in a different polyphrase, with their individual translations, and there will be another polyphrase, of "n-m" type, to contain the 2 complete segements.

The first problem we encounter with the available parallel corpora it is that there is no tool to visualize their contents at a glance, sentence by sentence, nor to show the fine correspondences between subsentential segments. In addition, in the case of UNL documents, we cannot visualize at the same time a sentences in several languages and its corresponding UNL graph. Lastly, it is not possible to see successive versions in parallel.

When it comes to evaluation, we can only see the monolingual files, and associated statistical measurements (NIST, BLEU...), but we can never confront them with the real translations and make a direct subjective evaluation.

2.3 Detailed objectives

The objectives of TraCorpEx project are as follows.

2.3.1 Construction of a software platform

We want to build an environment, which supports the import and the export of parallel corpora, the preparation of the data for automatic translators, the postedition (HAMT), the evaluation (various feedbacks methods) and finally a preparation of "feedbacks" to the developers of used MT systems.

2.3.2 Addition of new languages

Starting from parallel corpora, we want to add one or more languages (those of the Papillon project for the Tanaka corpus, French and Arabic for the BTEC corpus).

2.3.3 Evaluation of MT systems

We also wish that the same platform makes it possible to evaluate automatic translators with automatic methods such as NIST, BLEU, PER, and to use this possibility in CSTAR, to evaluate the Chinese-English and Japanese-English translations. To evaluate the results of various MT systems will also enable us to determine "the best" (or less bad!) translation, proposable to a contributor as a starting point for revision.

We also want to test a hypothesis by the second author: the quality of the translations could also be evaluated using calculations of distances between sentences and reverse translations.

2.3.4 Feedbacks to developers of MT systems

We also want to give feedbacks to the developers of the systems used (unknown words, badly translated sentences...), and a comparative presentation between the various translation systems.

The whole of the objectives of this project led us to propose interactive Web interfaces allowing us to chooses, use, compare, publish machine translations corresponding to several language pairs, and to contribute to the improvement of the results by sending feedbacks to the developers of these systems.

2.4 The PolyphraZ platform

PolyphraZ is a software platform making it possible at the same time to visualize the available corpora on the Web by showing several languages, with the choice of the user and to work on a basis of "polyphrases" initialised from these corpora while making it possible to control all functions described above (call of MT systems, distance computation, collaborative postedition, evaluation).

2.4.1 General architecture

We follow the software architecture of the Papillon platform.

We classify the objects to handle in three types

Raw corpus sources

Sources transformed into our XML format CXM. (Common Example Markup) and coded in UTF-8, for visualization "just as they are", then in CPXM format, DTD for parallel visualization.

MPM: multilingual polyphrase memory

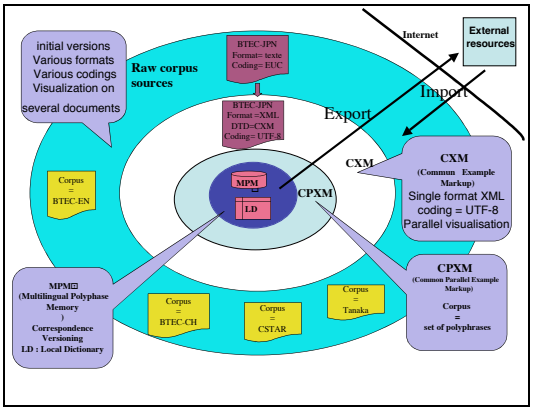


Figure 1: objects of the PolyphraZ platform

2.4.2 Intended users of PolyphraZ

We distinguish four principal users: the preparer, the reader ("normal" user), the posteditor and the manager.

- **The preparer**

  His role consists in calling translation systems, thereby parameterizing them as well as possible, which supposes a certain linguistic ability (to compare the results of various parameter settings, and of various segmentations in "blocks", each corresponding to some parameter settings).

  The preparer can also call objective evaluation methods (NIST, BLEU...) on the results of translation, tune with parameters to compute distances between sentences (results of translation and/or reverse translations), and post the results. The distance computation produces, in addition to a value, a XML string from which a “track changes” presentation can be generated. The preparer can also set the parameters determining "the best" suggestion among the various translation candidates.

- **The reader (normal user)**

  A reader can visualize the data (the original, various translations, and distances between the character strings) through Web interfaces, but is not allowed to edit the translations.

- **The translator-posteditor**

  The translator-posteditor is a contributor who translates from scratch or revises proposed translations (MT results or translations of similar sentences found in the MPM or in other TM put in CPXM or MPM format). There is an editable area to modify the active sentence. One can also ask for global modifications (ex: "SVP" changed into "s'il vous plait" in transcribed spoken utterances) and correct or supplement the local dictionary attached to the MPM. The system uses the reference sentences already produced like a translation memory. PolyphraZ is thus also a system of assistance to the translator, limited to the translation of sets of sentences (or titles), with less functionalities than commercial TWS, but usable for collaborative volunteer work by non-professionals.

- **The manager**

  The last type of user is the manager, who will produce from a MPM "feedbacks" for the developers of the MT systems used. A manager cans himself be a developer of an MT system.

  He can draw up a list of unknown words and words badly translated by each system (produced from the traces of distance computations). A second function is to propose for these words suggestions of translation from the "reference" translations obtained after human

revision. Finally, it is possibile to provide a presentation of the evaluations and comparisons between the results of the various systems used and/or their various parameter settings.

2.4.3 Implementation of PolyphraZ

Programmed in standard Java under the Enhydra development environment used for the dynamic and multilingual Papillon web site, PolyphraZ is multi-platform (MacOS-X/Unix/Linux, Windows).

2.5 Scenarios

The use of PolyphraZ can be divided in 3 parts: setting of the data under three different formats (CXM, CPXM, MPM).

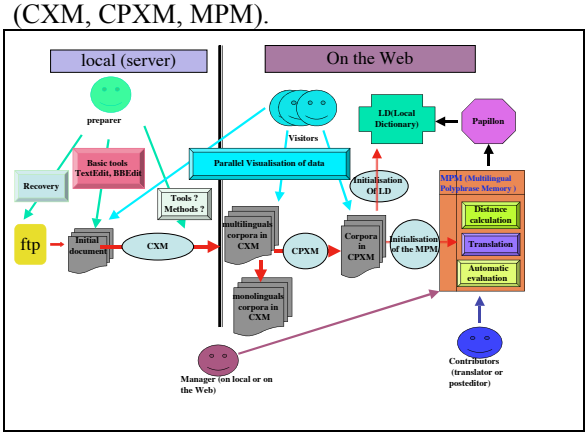


Figure 2 : scenarios for using PolyphraZ

2.5.1 CXM (Common eXample Markup)

In order to manipulate a single format (XML) and a single encoding (UTF-8), we automatically convert into the CXM format the imported data (corpus, text aligned...). CDM is defined in the same spirit as the CDM (Common Dictionnary Markup) of the Papillon project.

```
<?xml version="1.0" standalone="no" ?>
<!DOCTYPE document SYSTEM "CSTAR_BTEC_DTD.dtd" >
<document>
 <information documentname="CSTAR-corpus BTEC EJ"
 creation-date="Tue May 21 JST 2002"
 modification-date="Tue May 21 JST 2002"
 coding-set="UTF-8"
 number-of-language="2"
 number-of-sentences="162320"/>
 <sentence sentence-id="000001">
 <sentence xml:lang="EN">
 <segment segment-id="1">
Hamburger and stew on the right side and salad, please.
 </segement>
 </sentence>
 <sentence sentence-id="000001">
 <sentence xml:lang="IT">
 <segment segment-id="1">
Humburger e stufato dalla parte destra e insalata,per favore.
 </segement>
 </sentence>
 </document>
```

Figure 3: example XML file conforming to the CXM.dtd

2.5.2 CPXM.dtd (Common Parallel eXample Markup)

A second Java program transforms all CXM files corresponding to a given multilingual parallel corpus of sentences to the CPXM format (see appendix 2). In this format, we introduce the "polyphrase" XML element, which is a set of monolingual components, each containing possibly one or more proposals.

2.5.3 MPM.dtd (Multilingual Polyphrase Memory)

The MPM data structure is under construction. It is intended for the management of the correspondences between the various linguistic versions as well as the modifications which can be made, and to keep the history of the modified files. As shown in the following figure, a MPM of PolyphraZ can contain a set of versions and alternatives of the sentences, as well as the results of various computations.

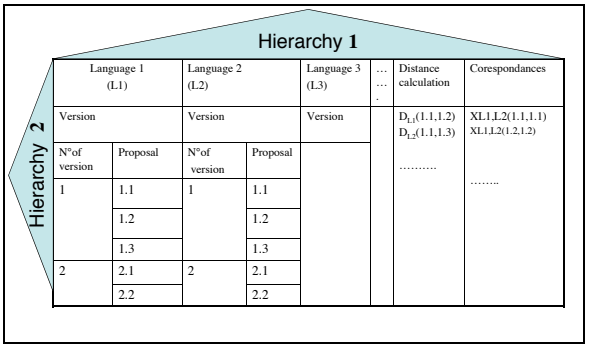


Figure 4 : logical view of a MPM

We give a first version of the MPM DTD in appendix 3.

2.5.4 Parallel visualization

PolyphraZ can visualize polyphrases in parallel from corpora in CPXM or MPM formats. This functionality is useful to compare translations, and is made available to readers; translators revisors, and managers.

Figure 5: parallel visualisation of the BTEC (extract)

2.6 Evaluation of translation results

We have programmed and integrad in PolyphraZ three evaluation methods (NIST, BLEU and distance calculation). NIST and BLEU are well known. Let us give more details about distance calculation between 2 sentences.

The distance we compute between two strings is a linear combination of two edit distances, one at the level of characters, the other at the level of words. In general, the edit distance between two strings P1 and P2 of atoms (characters or words here) is the minimal number of suppressions, insertions or replacements of atoms necessary to transform P1 into P2 or, equivalently, P2 into P1. To compute the edit distance between P1 and P2 at the level of words, one segments them into words, computes the character distances between words of P1 and words of P2, and then computes the word distance using words as "large characters".

We use the well-known dynamic programming algorithm of (Wagner, Fischer, 1974). To combine the two levels (characters and words), we use the formula:

$D = (\alpha D_{char} + \beta D_{word})/(\alpha+\beta)$; $\alpha + \beta = 1$

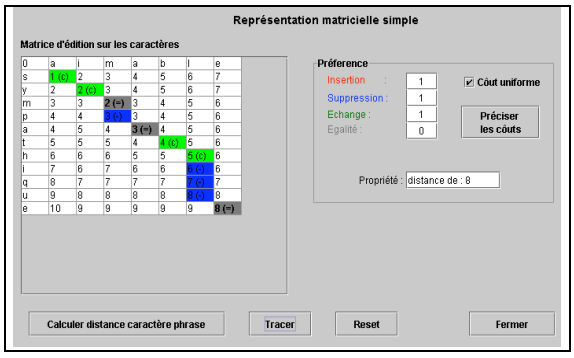


Figure 6:trace of the Wagner and Fischer algorithm

2.6.1 "Track changes" visualisation

This representation corresponds to the presentation used by Microsoft Word in "Track changes" mode. It is very readable. In certain cases, the representation at the level of the characters is more compact and readable that at the level of words, while it is the opposite in other cases. In fact, this

representation is not "faithful" to the trace, because a sequence of exchanges is transformed into a sequence of suppressions and a sequence of insertions.

~~ai~~sympa~~bl~~thique

Figure 7:"Track changes" display

One interesting and today unsolved problem is how to merge the 2 levels: given 2 sentences and their character and word edit distances, necessarily both minimal, how to produce a trace which would be "the best" or "a best" combination of the 2 traces?

2.6.2 Representation with 3 lines

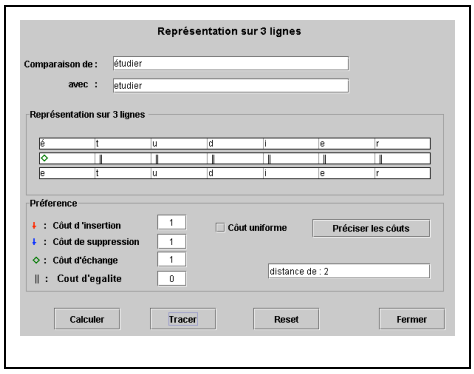


Figure 8 : 3 lines representation

This representation is simpler to understand, but takes more space.

◇ represents the exchange of a character by another,

|| represents the equality between two characters

↓ represent the suppression of the 1st character,

```
<?xml version="1.0" standalone="no" ?>
<!DOCTYPE document SYSTEM "CSTAR_BTEC_DTD.dtd" >
<document>
 <information documentname="CSTAR-corpus BTEC EJ"
 creation-date="Tue May 21 JST 2002"
 modification-date="Tue May 21 JST 2002"
 coding-set="UTF-8"
 number-of-language="2"
 number-of-sentences="162320"/>
 <sentence sentence-id="000001">
 <sentence xml:lang="EN">
 <segment segment-id="1">
 Hamburger and stew on the right side and salad, please.
 </segement>
 </sentence>
 <sentence sentence-id="000001">
 <sentence xml:lang="IT">
 <segment segment-id="1">
 Humburger e stufato dalla parte destra e insalata,per favore.
 </segement>
 </sentence>
 </document>
```

Figure 9 : XML representation

3 Conclusion

The CXM and CPXM levels of PolyphraZ are already used. They have allow us to import the BTEC multilingual corpus of parallel sentences (into the common CPM format), to transform it (163000 sentenes in 5 languages) into files in CPXM formats, and to visualize it[1] on the web. The Tanaka corpus should be available when this paper will be presented. The "inner" level of MPM (Multilingual Polyphrase Memory) is almost completed. It will also support versioning.

In the future, we plan to use MPMs not only to handle multilingual corpora of parallel sentences, but also like "pivots", to establish the sentence-level correspondence between parallel monolingual structured documents. If no high quality TWS (like Trados, TM2, Déjà Vu; Transit, etc.) is available, PolyphraZ could be used as a "bare bone" TWS, directly through the web, in the Montaigne[2] spirit.

We are also studying how to integrate into a MPM structure "generators" specifying classes of sentences (automata for messages with variables and variants, regular expressions for CSTAR IF expressions, etc.), and to use them to extend a MPM not only "in width" (addition of new languages), but also "in height", by the automatic creation of new "statements", natural and/or formal.

Références

A-B.Assimi (Assimi,2000). *Gestion de l'évolution non centralisée de documents parallèles multilingues*, Nouvelle thèse, UJF, Grenoble, 31/10/00, 200□.

A-B.Assimi & C.Boitet (Assimi&Boitet,2001) *Management of Non-Centralized Evolution of Parallel Multilingual Documents*. Proc. Internationalization Track, 10th International World Wide Web Conference, Hong Kong, May 1-5, 2001, 7 p.

Ch.Boitet (Boitet, 2003) *Approaches to enlarge bilingual corpora of example sentences to more languages*, Papillon-03 seminar, Sapporo, 3-5 July 2003 12□.

Ch .Boitet & Tsai W.-J (Boitet & W-J 2002). *Coedition to share text revision across languages. Proc. COLING-02 WS on MT*, Taipeh, 1/9/2002, 8 p.

[1] The full corpus is only accessible to members of CSTAR-III, so that we show only extracts corresponding to parts which are or will be published for the open evaluation of various MT systems to be presented at IWSLT-04.

[2] Mutualization Of Nomadic Translation Aids for Groups on the NEt (Mutualisation d'Outils Nomades de Traduction avec Aides Informatiques pour des Groupes sur le NEt).

H.Vo-trung (Vo-trung, 2004) *Réutilisation de traducteurs gratuits pour développer des systèmes multilingues*, accepté à la conférence RECITAL 2004, avril 2004, Fès, Maroc.

N.Hajlaoui, Ch .Boitet (Hajlaoui, Boitet, 2003a), *A "pivot" XML-based architecture for multilingual, multiversion documents□ parallel monolingual documents aligned through a central correspondence descriptor and possible use of UNL*, Convergences'03, Alexandria, 2-6 December 2003.

N.Hajlaoui, Ch.Boitet (Hajlaoui, Boitet, 2003b), *Modélisation de la production de phrases, projet franco-tunisien entre l'équipe* GETA, CLIPS, UJF, Grenoble et université de Sousse, Tunisie, 25 p.

N.Hajlaoui (2002) *Gestion des versions des composants électroniques virtuels*. Rapport de DEA, CSI, INPG, juin 2002, 80 p.

R.Wagner & Michael.Fischer (Wagner, Fischer ,1974) *The String-to-String Correction Problem* ACM Journal of the Association for Computing Machinery, Vol. 21, No 1, Janvier 1974.

W.-J.Tsai (Tsai,2001) SWIIVRE *a web site for the Initiation, Information, Validation, Research and Experimentation on UNL*. Proc. First UNL Open Conference - Building Global Knowledge with UNL, Suzhou, China, 18-20 Nov. 2001, 8 p.

(C-STAR-III) *C-STAR project*, http://www.c-star.org/

(Papillon) *Projet PAPILLON de construction coopérative d'une base lexicale multilingue et de construction de dictionnaires*, http://www.papillon-dictionary.org/

(TraCorpEx) projet TraCorpEx

http://www-clips.imag.fr/geta/User/najeh.hajlaoui/tracorpex/index.html

(UNL) *Universal Networking Langage (UNL) project*, http://www.undl.org/

Appendices

```
<!-- CXM.dtd  (Common eXample Markup ) is a
  DTD  which  describes the corpora
  (multilingual or monolingual), it is the
  simplest format for imported data.

$Author:  Najeh Hajlaoui
  najeh.hajlaoui@imag.fr
$Date: 2003/12/10 01:28:30 $ -->
<!ELEMENT document (information, sentence*) >
<!ELEMENT information (#PCDATA) >
<!ATTLIST information document-name  CDATA
  #REQUIRED>
<!ATTLIST information   creation-date
  CDATA #IMPLIED>
<!ATTLIST information modification-date
  CDATA #IMPLIED>
<!ATTLIST information   coding-set    CDATA
  #IMPLIED>
<!ATTLIST information number-of-languages
  CDATA     #IMPLIED>

<!ATTLIST information   number-of-sentences
  CDATA #IMPLIED>

<!ATTLIST sentence   sentence-id    CDATA
  #REQUIRED>
<!ATTLIST sentence xml:lang CDATA #REQUIRED>

<!ELEMENT sentence (segment*) >
<!ATTLIST segment   segment-id    CDATA
  #REQUIRED>
<!ELEMENT segment (#PCDATA) >

<!-- Document is a set of sentences, each
  sentence is defined
by an identifier called sentence-id and also
  by an attribute which indicates the
  language -->

<!-- number-of-languages is the total number
  of languages constituting the document; if
  the document is monolingual, number-of-
  languages =1   -->

<!-- number-of-sentences is the total number
  of sentences constituting the document -->

<!-- Each sentence is a set of one or more
  possible  segment; each segment is
  identified by an attribute called  segment-
  id -->
```

Appendix 1 : CXM.dtd (Common eXample Markup)

```
<!-- CPXM.dtd  (Common Parallel eXample
  Markup ) is a  DTD  which  describes the
  multilingual documents (m languages),
  multiversions (n versions) (n>m), it
  allows the description of a collection of
  polyphrases in a single format and
  encoding.
$Author:  Najeh Hajlaoui
  najeh.hajlaoui@imag.fr
$Date: 2003/06/10 01:28:30 $ -->
<!ELEMENT document (information,
  polyphrase*) >
<!ELEMENT information (#PCDATA) >
<!ATTLIST information document-name  CDATA
  #REQUIRED>
<!ATTLIST information   creation-date
  CDATA #IMPLIED>
<!ATTLIST information modification-date
  CDATA #IMPLIED>
<!ATTLIST information   coding-set
  CDATA #IMPLIED>
<!ATTLIST information number-of-languages
  CDATA                 #IMPLIED>
<!ATTLIST information number-of-
  polyphrases CDATA #IMPLIED>

<!ELEMENT polyphrase (monolingual-
  component*) >
<!ATTLIST polyphrase   polyphrase-id
  CDATA #REQUIRED>

<!ELEMENT  monolingual-component
  (segment*) >
<!ATTLIST monolingual-component xml:lang
  CDATA #REQUIRED>
<!ELEMENT  segment (proposal) >
<!ATTLIST proposal   proposal-id    CDATA
  #REQUIRED>
<!ELEMENT  proposal (#PCDATA) >

<!-- number-of-languages is the total
  number of languages appearing in the
  document; if the document is monolingual,
  number-of-languages =1   -->
<!-- number-of-polyphrases is the total
  number of polyphrases constituting the
  document -->
<!-- A polyphrase is a set of monolingual
  components, each containing 1 or more
  possible proposals. Every polyphrase is
  identified by a number called polyphrase-
  id -->
<!-- Each monolingual component is a set
  of one or more possible renderings of the
  segment in question; it is identified by
  an attribute which indicates the language
  -->
<!-- Segment represents the level of
  alignment, it is usually a sentence -->
```

Appendix 2 : CPXM.dtd (Common Parallel eXample Markup)

```
<!-- MPM.dtd  (Multilingual Polyphrases
  Memory ) is a  DTD  which  allows the
  generation of sentences aligned in
  several languages and the management of
  the correspondence between these
  sentences.
$Author:  Najeh Hajlaoui
  najeh.hajlaoui@imag.fr
$Date: 2003/01/28 21:28:30 $ -->
<!ELEMENT document (information,
  generator*, node-of-correspondence*) >
<!ELEMENT information (#PCDATA) >
<!ATTLIST information document-name
  CDATA #REQUIRED>
<!ATTLIST information   creation-date
  CDATA #IMPLIED>
<!ATTLIST information modification-date
  CDATA #IMPLIED>
<!ATTLIST information   coding-set
  CDATA #IMPLIED>
<!ATTLIST information number-of-languages
  CDATA #IMPLIED>
<!ATTLIST information number-of-generator
  CDATA #IMPLIED>

<!ELEMENT generator (instance*) >
<!ATTLIST generator   original    CDATA
  #REQUIRED>
<!ATTLIST generator   context    CDATA
  #REQUIRED>

<!ELEMENT  instance (segment*) >
<!ATTLIST instance xml:lang CDATA
  #REQUIRED>
<!ATTLIST segment node-of-corespondance-
  id CDATA #REQUIRED>
<!ELEMENT  segment (proposal) >
<!ELEMENT  proposal (#PCDATA) >

<!-- number-of-languages is the total
  number of languages appearing in the
  document; if the document is
  monolingual, number-of-languages = 1 -->
<!-- number-of-generator is the total
  number of generator appearing in the
  document -->
<!-- A generator is a set of original
  sentences and  their instance  -->
<!-- A instance is a set of one or more
  possible renderings of the segment in
  question; it is identified by an
  attribute which indicates the language
  -->
<!-- Segment represents the level of
  alignment, it is usually a sentence -->
<!-- A node-of-correspondence-id
  represents the link of corespondance
  between the diférents proposals of
  translation -->
```

Appendix 3 : MPM.dtd (Multilingual Polyphrase Memory)

Multilingual Text Induced Spelling Correction

Martin REYNAERT
Induction of Linguistic Knowledge, Computational Linguistics and AI, Tilburg University
Warandelaan 2, 5000 LE Tilburg,
The Netherlands,
reynaert@uvt.nl

Abstract

We present TISC, a multilingual, language-independent and context-sensitive spelling checking and correction system designed to facilitate the automatic removal of non-word spelling errors in large corpora. Its lexicon is derived from raw text corpora, without supervision, and contains word unigrams and word bigrams. The system employs input context and lexicon evidence to automatically propose a limited number of ranked correction candidates. We describe the implemented trilingual (Dutch, English, French) prototype and evaluate it on English and Dutch text, monolingual and mixed, containing real-world errors in context.

1 Introduction

The EAGLES final report on 'Evaluation of Natural Language Processing Systems' lists as a 'dream tool' (EAGLES-I, 1996):

> A multilingual spelling checker which automatically recognizes what language is being dealt with and switches to the appropriate spelling checker for that language.

Our Text Induced Spelling Correction algorithm (TISC) represents such a tool in its current three language version, but we explore the possibility of *not* performing explicit language detection. This was prompted by the observation that language detection in an isolated-word system may easily get confused. Take the recent Dutch newspaper Metro headline 'Crime passionel in Gronings zwembad' [Crime of passion in Groninger swimming-pool (21-10-2003)], which is a typical example of mixed language text, containing a typo **passionel*, which in French should be spelled *passionnel*. The Microsoft Proofing Tools (MPT), for instance, can be set to automatically detect the language. Given the journalist is Dutch, it would typically have Dutch as its default language and so will not switch languages given the headline's first word *crime* is present in the Dutch dictionary, too. It will then encounter **passionel* and propose the correct, Dutch, forms: *passionele* and its lemma *passioneel*. Whereupon the journalist, not being too versant in French, is likely to let his original pseudo-French **passionel* stand. The Dutch part of the web provides many more instances of this same error, as does the English, for that matter. Our system being context-sensitive, we therefore explore whether its word bigrams alone aid the detection and correction of this kind of error, even when no further explicit language detection is done and no switching to another language dictionary occurs, its dictionary containing a mix of its various languages. In order to present our findings, we first describe our novel correction mechanism (section 2), explain how we effect detection in light of a noisy lexicon (section 4), derived from one or more language corpora (section 3) and present the evaluation results obtained on Dutch, English and mixed Dutch-English language texts (section 5).

2 The correction algorithm

We develop the idea of using the corpus itself as the basis on which to build a spelling correction system.

2.1 Anagram hashing

We line up all those word forms present in the corpus that consist of the same set of characters and use that as the basis for a corpus-derived lexicon. A means to do this in a completely unsupervised way was found in the theory of hashing, be it in the 'bad' part of it, in the normally avoided generation of collisions. Hashing has before been applied to spelling checking (Kukich, 1992), but we know of no prior work based on hash collisions. Collisions occur when

Anagram key	anagrams
75123219269	gerti, giert, griet, regit, riget, tiger, tigre
95176774701	ce tigre
95666874202	de griet, de tigre, dreig te, giert de, tigre de
107081254058	dreigt u, du tigre, it urged, u dertig, u dreigt, urged it
115780446077	de gierst, de tigres, gerst die, get rides, griste de, its greed, tigres de
127194825933	de rustig, drug ties, it surged, rustig de, surgit de, tigres du, urged its
129962785833	a stringed, and tigers, art design, dangers it, de ratings, de ratings, drang iets, gradins et, grand site, granted is, gratin des, is granted, its danger, its garden, rating des, ratings de, red giants, sign trade, tigers and, tigre dans

Table 1: Extract from a trilingual (English, Dutch, French) TISC lexicon with the anagram keys and associated, chained anagrams

the mathematical function used to bin the information, puts more than one item of information in a single bin (Knuth, 1981). The mathematically simple function introduced and exploited here does precisely that, for all strings containing the precise same set of characters.

So, for each word type or word type combination (compound or word bigram) to be included in the TISC lexicon, we obtain a numerical value, which will serve as the hash key. The formula represents the mathematical function we devised to do this, where f is a particular numerical value assigned to each character in the alphabet and c_1 to $c_{|w|}$ the actual characters in the input string w.

$$Key(w) = \sum_{i=1}^{|w|} f(c_i)^n$$

In practice, we use the ISO Latin-1 code value of each character in the string raised to a power n. We currently use 5 as the value for n. This was empirically derived, lower values do not produce collisions between anagrams only. The rather large natural number produced by this function in effect inflates the difference between any two characters to such a degree, that all strings containing the same set of characters receive the same natural number. This means that all anagrams, words consisting of a particular set of characters and present in the lexicon, will be identified through their common numerical value. So, in that the collisions produced by this function identify anagrams, we refer to this as an *anagram hash* and to the numerical values obtained as the *anagram keys*.

In the implementation the anagram keys and their associated word forms are stored in a regular hash. The anagram key will enable us to look up immediately whether any string consisting of the same character set as the input string was encountered in the corpus. When not present in the lexicon, close (numerical) neighbours might very well be present, and simple arithmetic will allow us to identify and retrieve these. This representation makes the implementation computationally tractable. The net effect of obtaining anagram hash key values is that it provides a cheap abstraction from the surface sequence of characters which further allows, through simple addition, subtraction or both, for moving from one particular combination of characters to another. The numerical difference between e.g. any verb possibly ending in *-ise* or *-ize* will always be the same. Subtracting the anagram key value of the *s*-variant from the anagram value of the *z*-variant will produce the same numerical result for all these pairs as does subtracting the anagram value for the single character *s* from the anagram value for *z*, namely: $z = 122^5 = 27,027,081,632$ and $s = 115^5 = 20,113,571,875$, difference: $27,027,081,632 - 20,113,571,875 = 6,913,509,757$. The numerical difference between e.g. *randomize* and *randomise* equals $136,483,404,939 - 129,569,895,182 = 6,913,509,757$. The same goes for all systematic spelling variations between e.g. American and British English or in probably any other alphabetic language.

2.2 Anagram key based correction

Anagram key based spelling correction is an inexpensive solution to the string correction problem as it does not entail expensive searching: it uses the non-search strategy implied in hashing. Based on a word form's anagram key it becomes possible to systematically query the lexicon for any variants present, be they morphological, typographical or orthographical. These variants can all be seen as variations of the usual taxonomy in terms of *trasnpositions, *deletons, *inserrtions or *substatutions (Damerau, 1964).

transpositions These we get for free: they have the same anagram key value, so when queried, the lexicon returns the correct form and its anagrams (if any).

deletions We iterate over the alphabet and query the lexicon for the input word anagram value plus each value from the alphabet.

insertions We iterate over the list of anagram values for the character unigrams and bigrams collected from the input type and

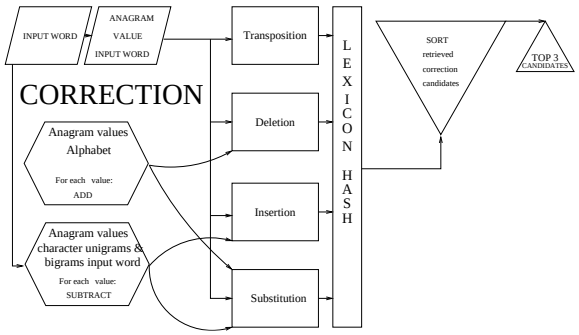


Figure 1: The correction module

query the lexicon for the input word anagram value minus each of these values.

substitutions We iterate over both lists adding each value from the one and subtracting each value of the second to the input word anagram value and repeatedly query the lexicon.

We thus retrieve all numerical near-neighbours (NNNs) from the lexicon and apply standard string matching techniques to retain those that either in front or back match the input type for a specific amount of characters, depending on the input type's length. After doing so, we iterate over the list of NNNs retained and upgrade the actual retrieval counts for those that have the greater substring matches and whose Levenshtein distance (LD)(Levenshtein, 1965) does not exceed 4 (the algorithm is not in itself limited to a particular LD). The elements of this list have thereby been ranked and the top n are then proposed as correction candidates. This ranking is an automatic side effect of the algorithm which produces more hits on the actual nearest NNN's. A deletion error, e.g. such as **cateory*, will return the correct 'category' on the basis of adding the anagram value for 'g' as well as of substituting the value for 'e' with that for 'eg' and substituting the value for 'o' with that for 'go'. The redundancy inherent to our algorithm thereby produces the desirable side-effect of converging on what is usually the best correction candidate by returning it more often than less likely candidates.

3 TISC corpus-derived components

3.1 The Lexicon

The English corpus we used was the New York Times (1994-2002) material available in the LDC Gigaword Corpus (NYT) (Graff, 2003). For Dutch we used both the ILK Corpus[1] and the Twente Corpus[2] (TWC). For French we used 8 years ('91-'98) of Roularta Magazines[3]. Statistics on these corpora are presented in table 2.

Corpus	NYT	ILK-TWC	ROULARTA
language	English	Dutch	French
tokens	1,106,376,695	681,686,340	52,722,253
bigrams	11,246,986	9,927,378	1,270,600
unigrams	672,502	861,604	144,943
keys/anagr.	10,287,826	9,000,131	794,308

Table 2: Statistics of NYT, ILK-TWENTE and ROULARTA corpora and lexicons. Bigrams with frq>2. Unigrams derived from these. French key-anagram ratio based on frq>4.

A TISC lexicon is derived from a large corpus of tokenised, but otherwise raw text, from which all XML or other tags have been discarded. We normalise the corpus by replacing all word-external punctuation by a single unique mark, as well as all digits and numbers by another. We apply a rule-based tokenizer and use the CMU Statistical Toolkit for deriving a bigram frequency list from the corpus (Clarkson and Rosenfeld, 1997). We discard the tail of the bigram list below a given threshold frequency, partly to ensure we do not incorporate the bulk of erroneous types present in the corpus. The effect of varying the threshold frequency is discussed in (Reynaert, 2004).

To make a multilingual version we concatenate the different languages' bigram lists at this point. Next the frequency information is discarded and a unigram list derived from the retained part of the bigram list. We lowercase the unigram list and concatenate the three lists obtained, removing any doubles. We finally compute the anagram key values for the unigram/bigram list. Together, the anagram keys and their lined-up unigrams or bigrams constitute the lexicon. Note that the lexicon will contain names and higher frequency errors.

3.2 The alphabet

Transformations on the word type to be evaluated are necessary in order to identify correction candidates. These transformations occur on the anagram key of the word type under consideration on the basis of numerical, i.e. anagram, values for the alphabet used, which are

[1] http://ilk.uvt.nl/ilkcorpus/

[2] http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html

[3] http://www.roularta.be/en/products/default.htm

read in at the start of run time. Our alphabet consists of the anagram key values for all character unigrams (e.g. a = 8,587,340,257, Z = 5,904,900,000) and character bigrams (e.g. ab = ba = 17,626,548,225, $i\ddot{e}$ = $\ddot{e}i$ = 729,465,962,500) we want to work with. The list currently contains 442 anagram values. These have been derived from character unigram and character bigram counts on the corpus.

3.3 The cooccurrence information

From the word bigram and unigram lists we derive cooccurrence information for all the word types present. For each word type we count the number of times it forms the:

- left part of a compound (LPC)
- right part of a compound (RPC)
- left part of a bigram (LPB)
- right part of a bigram (RPB)

Note that these cooccurrence counts (COOC) are counts on word types and not on word tokens. The COOC table contains only the counts per word-type, not the actual cooccurring word types.

4 TISC: the implementation

4.1 Zipf filters

Recall that Zipf stated that the frequency of a word is inversely proportional to its length (Zipf, 1935). This implies that we should expect to see more combinations of any given short word than of longer words. A long compound, e.g. one composed of three or more shorter words, cannot reasonably be expected to combine with very many more words. Short words can be expected to combine in a myriad of ways, be it as part of compounds or of numerous bigrams. It is this idea we exploit in what we would like to call the *Zipf Filters* implemented in our prototype. We make the number of expected cooccurrences of a word dependent on the length of the word form. This then allows to detect anomalies in the COOCs for particular word types. We posit a particular amount of times a string or substring is seen as sufficient to conclude the string is likely well-formed as it is highly productive. To this end we take a constant, which is higher for the shorter strings and lower beyond a particular amount of characters, divided by the number of characters in the string, or the string's length. We compare the COOCs of a string to be evaluated with the outcome of this calculation and accept the string as being well-formed when the COOCs are higher, reject and thus send on to the correction module, when lower.

4.2 Compound splitting

Given that a language such as Dutch to a large degree allows for compounding, any text may contain quite a number of previously unseen compounds. While iterating over the input word string to compute its anagram value, TISC repeatedly queries the lexicon to check for the presence of the substring handled so far. If this is successful for the string as a whole, the substrings, if any, which show the best balance between length and COOCs are stored with their anagram values. If no full parse was possible, the process is repeated from right to left and a decision made over both the left-right and right-left parses and the split deemed most usable stored. TISC proposes a single particular split to be further provided to the checking and correction modules. The implementation currently allows for only a split in a left and right part.

4.3 Checking

The input text is first fully analysed: anagram values are added to the type list, frequencies of types and their compounding parts tallied, track kept of how many times the type was capitalised, recurrent LPC's not in the lexicon stored. Then, all the types are sent to the spelling checking module. Since we cannot content ourselves with simply checking whether a type is present in the dictionary or not, we query the cooccurrence information table to see whether the particular type's COOCs conform to our expectation of how many times a type of the given length should have been incorporated in the lexicon, i.e. the expectancy level or threshold set by the Zipf filter. If this is the case, the type is not further evaluated, which we will refer to as 'let go' . If not, the COOCs for its LPC and the RPC are evaluated against the threshold. We do not, at this stage, want to risk to lose too many of the erroneous types, so the level of expectancy is set rather high. We simultaneously check whether perhaps the lexicon contains possible bigrams based on the type's anagram key value with the value for a space added. All the types which did not conform to the expected levels or were found to be present with an additional space, are further evaluated. Further checks are:

- extra-space cases: If it turns out the lexicon contains only the inverted form with

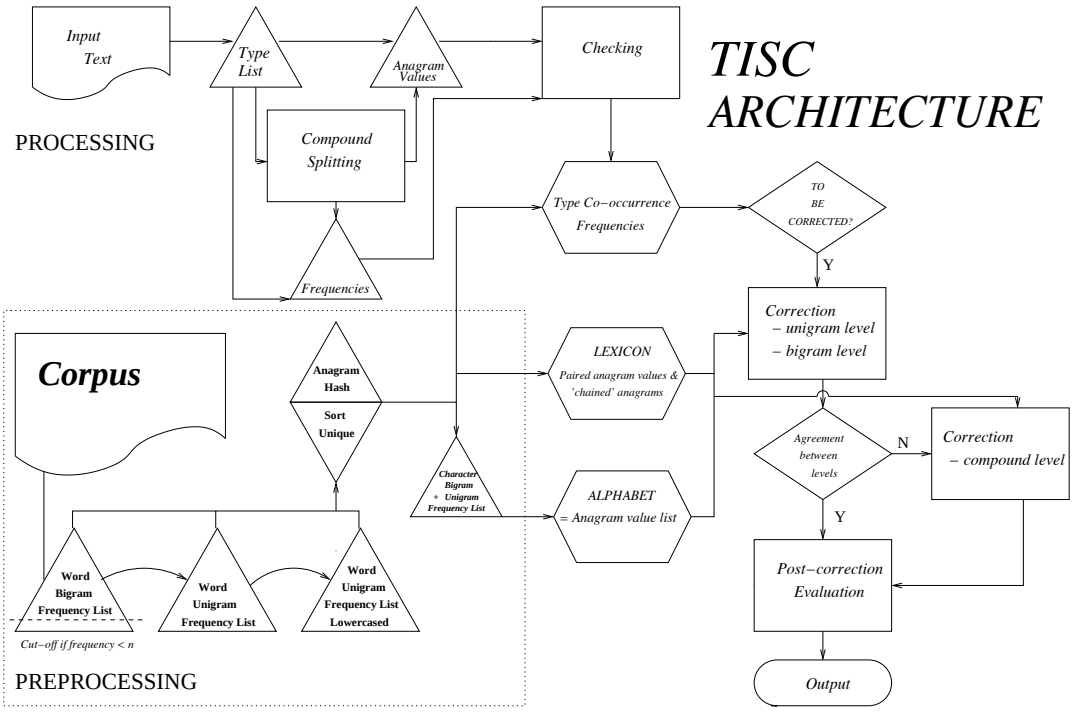


Figure 2: TISC's architecture

the added space (e.g. 'koffiebekertje' [coffee cup]: not in the lexicon, but 'bekertje koffie' [cup of coffee] is present), we accept the form as being correct, the rest are further evaluated.

- whether perhaps the LPC was seen in various other input text compounds or whether the RPC was perhaps seen as a word in its own right with a given frequency in the input text, the other part's COOCs conforming. Again those passing this test are let go.
- whether perhaps the COOCs for the LPC with first or all characters upper-cased conform to expectance.
- if the input type contains a dash, we check whether the COOCs for the type without the dash conform. Or perhaps whether the type without the dash but with an extra space is present in the lexicon.
- finally we check those forms for which the cooccurrence table contains no information at all. If the COOCs for their LPC and RPC exceed a high expectancy threshold, these are let go too.

All types not let go by one of these checks are sent on to the correction module.

4.4 Correction

By default, TISC's correction works on two levels, a third being invoked when these do not return satisfactory results. The unigram level consists of two tiers: unigram correction on the basis of the lexicon and on the basis of the list of input context derived types and compounding parts (with frequency threshold). On the bigram level, TISC performs context-dependent error correction, to some extent. It examines the 4 bigrams contained within a 2-1-2 window around the type in the input text (e.g. the green **bottel* was empty → the **bottel*, green **bottel*, **bottel* was, **bottel* empty). The only difference with the unigram correction module lies in the fact that for the 4 bigrams sent through the correction loop, all the correction candidates retrieved are stored in the same list. This produces more reliable counts after upgrading. After correction on these levels, the output candidates are compared and if both levels concur, i.e. the same candidate(s) were returned, they are accepted if they differ from the input type, or rejected (and 'let go') if not. When no output is returned by the unigram and bigram correction levels, or the results of these do not concur, the type is further checked on the third level, that of its substrings, i.e. the compounding parts returned by the compound

splitter. The compound correction level treats both LPC and RPC as words in their own right, queries the system for correction candidates in the same way as on the unigram level for both parts and finally concatenates the top candidates returned and proposes these as correction candidates. Given a sufficiently high frequency in the input text of the correct form for an incorrect compounding part, this may enable the system to correct the error even if the correct form is not present in the lexicon.

5 Evaluation

5.1 Evaluation method rationale

TISC ought to be compared to other context-sensitive spelling checking and correction systems applied to the task of detecting and correcting non-word errors. Alas, we know of none that have been evaluated on both detection and correction.

Brill and Moore have developed and evaluated an improved noisy channel-based correction system equipped with a language model, therefore context-sensitive, and reported state-of-the-art correction performance (Brill and Moore, 2000). They trained the system on 8,000 erroneous word forms. The system was given another 2000 erroneous word forms to correct under perfect conditions: all correct forms were present in the dictionary. They report an accuracy of 98.8% on the 3-best ranked correction candidates. We think this really constitutes the upper bound their system can reach, rather than its true accuracy. We get no idea of how this system would perform, if it were given both correct and incorrect words not available in the dictionary. In order to evaluate our system in the same way and in order for results to be comparable, we would have to be able to use the same 2,000 error list. This list does not seem to be available.

We therefore tried to next best thing, which is to try and see how an isolated-word spelling checking and correction system, which can easily be equipped with the same bi- and trilingual dictionaries as TISC, performs. ISPELL fulfils these requirements. Unfortunately, it does not perform ranking of the correction candidates. Either it sorts them alphabetically or not. This precludes reporting ranking scores here.

5.2 Test settings

We compare our results with those obtained by ISPELL (version 3.2.06) and MPT (version in Microsoft Office 2000, 9.0.3821 SR-1), as far as possible. For TISC and our trilingual version of ISPELL we varied the threshold at which the corpora's bigram lists were truncated (Frequencies: 4–10, 15, 20, 30, 40, 50 and 100). The TISC implementation used was the same for all tests as it contains no provisions specific to a particular language. For the monolingual tests both ISPELL and MPT were run with their standard US and standard Dutch dictionaries, the first in batch mode, the second manually emulating ISPELL's output for automatic evaluation purposes. For the multilingual test, we declined testing MPT's automatic language detection mode on the 145,100 token file. For both ISPELL and MPT we report the averaged scores of the three monolingual tests in contrast to the trilingual ISPELL and bi- and trilingual TISC test results.

5.3 Composition of the evaluation files

Statistics on the evaluation files are presented in table 3.

Dutch: For evaluation purposes, we proofread the Dutch version of the newspaper *Metro* and collected the non-word errors encountered (typically 0–4 a day). These were extracted from the online version[4] with the full article they appeared in. We used the first batch (Metro1) for development purposes. The second, similar, batch we reserved for testing purposes only (Metro2).

English: We manually collected 1093 erroneous types from the alphabetically sorted unigram frequency list of the Reuters Corpus (Lewis et al., 2003). We then extracted their contexts from the tokenized corpus. The context ran to the paragraph containing the error, as well as the paragraphs preceding and following it. We proofread these manually, which yielded another 105 errors. A preliminary Ispell run finally yielded another 24 overlooked errors. We ran our evaluations with these 1222 known errors. Statistics on the evaluation file are presented in table 3.

Dutch-English: For the bilingual tests, we concatenated both Metro files and the Reuters file and sorted the lines alphabetically, thereby obtaining a mixed language file.

5.4 Scoring and evaluation results

We measure performance in terms of the F-score. Given that the systems are presented

[4] `http://www.metropoint.com/cgi-bin/WebObjects/Metropoint.woa/wa/default`

	Metro1	Metro2	Reuters	Mixed
context	article	article	3 par.	mix
tokens	21,919	25,750	97,432	145,100
types	5,747	6,441	15,341	24,795
errors	129	123	1,222	1,474
error/type	2.25%	1.9%	8%	5.9%

Table 3: Statistics of the evaluation files

	Rec.	Prec.	F	frq
Dutch:				
MPT	0.66	0.1	0.17	-
ISPELL	0.60	0.07	0.12	-
TISC	**0.67**	0.60	**0.63**	5
TISC-BI	0.64	**0.61**	0.62	5
TISC-TRI	0.64	**0.61**	0.62	5
English:				
MPT	**0.94**	0.38	0.54	-
ISPELL	0.85	0.27	0.41	-
TISC	0.85	0.80	**0.82**	5
TISC-BI	0.81	**0.83**	**0.82**	4
TISC-TRI	0.84	0.81	**0.82**	5
Dutch-English:				
MPT-AVERAGE	0.74	0.19	0.3	-
ISPELL-AVERAGE	0.7	0.14	0.22	-
ISPELL-TRI	0.77	0.59	0.67	6
TISC-BI	**0.80**	0.77	0.78	6
TISC-TRI	0.79	**0.78**	**0.79**	5
D-E Upper bounds				
ISPELL-TRI-UPPER	0.84	0.63	0.72	5
TISC-TRI-UPPER	0.84	0.80	0.82	5

Table 4: Statistics of best test scores

with errors in a context, we do not solely measure their ability to correct incorrect forms (i.e. their accuracy), but also to discern between correct and incorrect input forms. Of the word forms for which correction candidates are returned, we check if the output contains the correct form. If so, the score for successful correction (recall) is augmented by one, no account being taken of the ranking of the correction candidates, because ISPELL does not have a ranking mechanism. For all the forms marked by ISPELL or MPT as 'not in the dictionary' the score for false positives (precision errors) is incremented by one. The same goes for those forms for which the systems return correction candidates, but where the correct one is missing. The results presented in table 4 were obtained on the word types, for all systems.

5.5 Discussion

Monolingual task: For both languages, TISC's lower thresholded lexicons consistently produce the highest precision. Recall rises as the threshold is set higher, to drop again, as does precision, with more and more information not being available. More context causes precision to drop: more words to be checked create more opportunity to report false positives. This is clearly demonstrated by the Dutch results, where the evaluation files contain a lower error to type ratio than the English one. The drop in precision given more context seems to us to be the main cause of current spelling checking systems not being able to attain automatic correction levels of performance, i.e. a level of precision where more errors would be removed than correct words erroneously replaced. The drop in recall for Dutch is certainly a result of its greater morphological diversity.

Bilingual task: Table 4 presents the best results on the bilingual English-Dutch correction task obtained by TISC and ISPELL with dictionaries based on the same bilingual (D-E) (BI) and trilingual (D-E-F) (TRI) bigram lists. These results are contrasted to the average of the monolingual results on the three evaluation sets obtained by ISPELL and MPT. A rather striking result is that ISPELL's performance is drastically improved by providing it with a much larger dictionary. The presence of names alone in the dictionary provided by us must account for the better part of the gain in precision.

We determined the upper bound for both trilingual systems by removing the errors present in the evaluation files from the bigram lists from which the lexicons were derived. Remember that the evaluation files were obtained from disjoint corpora, a number of these errors are therefore recurrent and may obtain relatively high frequencies. It can be seen that ISPELL with its simple dictionary look-up strategy is more sensitive to these than is TISC. This is a clear indication that TISC's error detection strategy based on COOCs and thresholds set by the Zipf filters works. TISC's main gain is due to its context-awareness and to its greater reach in terms of LD covered. So it corrects errors that are beyond ISPELL's scope, but still misses highly recurrent ones.

Simply mixing three languages seems to have no adverse effect on both TISC and ISPELL's capabilities of performing correction to these levels of performance. Nevertheless, the fact remains that this strategy entails that one particular type of errors will go undetected, namely those errors in a specific language that result in a valid word in one of the other languages in

this type of multilingual system. These would have to be called *bilingual or translingual confusables*. Our evaluation files happened to contain a few of them, e.g. polite which should have read 'politie' [police] in the Dutch evaluation set. The fact that these are a lot rarer than errors which do not form a valid word in any of the languages, obscures their effect. Note that these would throw a non-context-aware system which does attempt to do language detection off balance. We think context-awareness here too should help remedy this shortcoming of our non-language-detecting approach. Provided the error detection module is made to take into account the word bigram information in much the same way as the error correction module currently does, it should also be possible to detect these anomalies. And this may be a nice pointer to the way we should direct our future work, in that this at least hints at ways the harder task of detecting and remedying monolingual confusables (Kukich, 1992) may be tackled.

As a final note, we want to draw due attention to the fact, not overly stressed in the above, that we have developed a competitive spelling checking and correction system using nothing besides electronically available collections of text. For Dutch and English, of course, a great deal of natural language processing resources are available. We have deliberately ignored these, as there are a great many languages in this world for which little or no such resources have as yet been developed. The inexpensive approach outlined here, we hope, may help to remedy that.

6 Conclusion

We have presented TISC, a new algorithm for spelling checking and correction. We have outlined how the system is built up from large corpora of raw text. We have introduced a novel representation for lexical information which allows for an exact calculation of the difference between two character strings. Not only does this make the problem computationally tractable, it also allows for building a scaled system. We have shown that incorporating word bigrams, cooccurrence information about individual word types and context information derived from the input text, all combine to make multilingual spelling correction a competitive possibility. We have compared TISC with two state-of-the-art systems and shown that it outperforms both.

Acknowledgements

Heartfelt thanks to my supervisors Prof. Dr. Walter Daelemans and Dr. Antal van den Bosch for their trust and support, as well as to Dr. Sabine Buchholz for providing the tokenizer. This work was funded by the Netherlands Organisation for Scientific Research (NWO/FWO VNC 205-41-119).

References

E. Brill and R.C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proc. of the 38th Annual Meeting of the ACL*, pages 286–293.

P.R. Clarkson and R. Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings ESCA Eurospeech 1997*.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, Volume 7, Issue 3 (March 1964):171 – 176.

EAGLES-I. 1996. Final Report. In *Evaluation of Natural Language Processing Systems*, volume EAGLES DOCUMENT EAG-EWG-PR.2.

David Graff. 2003. The New York Times Newswire Service. *English Gigaword LDC-2003T05*.

Donald E. Knuth, 1981. *Sorting and Searching*, volume 2 of *The Art of Computer Programming*, section 6.4, pages 513–558. Addison-Wesley, Reading, Massachusetts, second edition.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

V.I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. In *Cybernetics and Control Theory*, volume 10(8), pages 707–710. Original in: Doklady Nauk SSSR 163(4): 845–848 (1965).

D. Lewis, Y.Yang, T.G. Rose, and F. Li. 2003. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*.

Martin Reynaert. 2004. Text induced spelling correction. In *Proceedings COLING 2004, Geneva*.

George Kingsley Zipf. 1935. *The psycho-biology of language: an introduction to dynamic philology*. The M.I.T. Press, Cambridge, MA, 1965 - 2nd. edition.

Authors Index