Ontology driven content extraction using interlingual annotation of texts in the OMNIA project

Achille Falaise, David Rouquet, Didier Schwab, Hervé Blanchon, Christian Boitet
LIG-GETALP, University of Grenoble
{Firstname}.{Lastname}@imag.fr

Abstract

OMNIA is an on-going project that aims to retrieve images accompanied with multilingual texts. In this paper, we propose a generic method (language and domain independent) to extract conceptual information from such texts and spontaneous user requests. First, texts are labelled with interlingual annotation, then a generic extractor taking a domain ontology as a parameter extract relevant conceptual information. Implementation is also presented with a first experiment and preliminary results.

1 Introduction

The OMNIA project (Luca Marchesotti et al., 2010) aims to retrieve images that are described with multilingual free companion texts (captions, comments, etc.) in large Web datasets. Images are first classified with formal descriptors in a lightweight ontology using automatic textual and visual analysis. Then, users may express spontaneous queries in their mother tongue to retrieve images. In order to build both formal descriptors and queries for the ontology, a content extraction in multilingual texts is required.

Multilingual content extraction does not imply translation. It has been shown in (Daoud, 2006) that annotating words or chunks with interlingual lexemes is a valid approach to initiate a content extraction. We thus skip syntactical analysis, an expensive and low quality process, and get language-independent data early in our flow, allowing further treatments to be language-independent. We use the lightweight ontology for image classifications as the formal knowledge representation tha determines relevant information to extract. This ontology is considered as a domain parameter for the content extractor.

We are testing this method on a database provided for the image retrieval challenge CLEF09 by the Belgium press agency Belga. The database contains 500K images with free companion texts of about 50 words (about 25M words in total). The texts in the database are in English only, and we ”simulate” multilinguism with partially post-edited machine translation.

The rest of the paper is organized as follow. We first depict our general architecture deployed for CLIA and then detail the various processes involved : interlingual annotation, conceptual vector based disambiguation and ontology driven content extraction. We conclude with the first results of experimentations on the CLEF09 data.

2 General architecture

2.1 General process

In our scenario, there are two types of textual data to deal with : companion texts in the database (captions), but also user requests. The two are processed in a very similar way.

The general architecture is depicted in figure 1. The main components, that will be described in detail, may be summarized as follows:

- Texts (both companions and requests) are first lemmatised with a language-dependent piece of software. Ambiguities are preserved in a Q-graph structure presented in section 3.1.2.

Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010, pages 52–60,
Beijing, August 2010

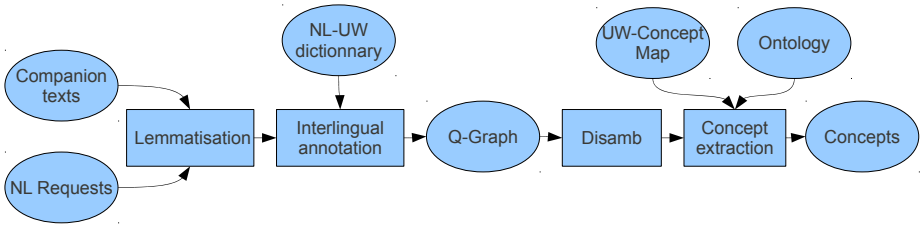

Figure 1: General architecture of CLIA in the OMNIA project

- Then, the lemmatised texts are annotated with interlingual (ideally unambiguous) lexemes, namely Universal Words (UW) presented in section 3.1.1. This adds a lot of ambiguities to the structure, as an actual lemma may refer to several semantically different lexemes.
- The possible meanings for lemmas are then weighted in the Q-graph through a disambiguation process.
- Finally, relevant conceptual information is extracted using an alignment between a domain ontology and the interlingual lexemes.

The conceptual information in the output may adopt different shapes, such as a weighted conceptual vector, statements in the A-Box of the ontology or annotations in the original text, etc.

In the case of OMNIA, conceptual information extracted from companion texts is stored in a database, while conceptual information extracted from users requests are transformed into formal requests for the database (such as SQL, SPARQL, etc.).

2.2 Implementation

The general process is implemented following a Service Oriented Architecture (SOA). Each part of the process corresponds to a service.

This allowed us to reuse part of existing resources developed on heterogeneous platforms using web interfaces (in the best case REST interfaces (Fielding, 2000), but frequently only HTML form-based interfaces). A service supervisor has been built to deal with such an heterogeneity and address normalization issues (e.g. line-breaks, encoding, identification, cookies, page forwarding, etc.).

This architecture is able to process multiple tasks concurrently, allowing to deal with users requests in real time while processing companion texts in the background.

3 Interlingual annotation

We present in this section the preliminary treatments of multilingual texts (image companion texts or user requests) that are required for our content extraction process (Rouquet and Nguyen, 2009a).

In order to allow a content extraction in multilingual texts, we propose to represent texts with the internal formalism of the Q-Systems and to annotate chunks with UNL interlingual lexemes (UW) . Roughly, we are making an interlingual lemmatisation, containing more information than simple tagging, that is not currently proposed by any lemmatisation software.

3.1 Resources and data structures

3.1.1 The Universal Network Language

UNL (Boitet et al., 2009; Uchida Hiroshi et al., 2009) is a pivot language that represents the meaning of a sentence with a semantic abstract structure (an hyper-graph) of an equivalent English sentence.

The vocabulary of UNL consists in a set of Universal Words (UW). An UW consists of:

1. a *headword*, if possible derived from English, that can be a word, initials, an expression or even an entire sentence. It is a label for the concepts it represents in its original language ;
2. a *list of restrictions* that aims to precisely specify the concept the UW refers to. Restrictions are semantic relations with other

UW. The most used is the “icl” relation that points to a more general UW.

Examples :

- book(icl>do, agt>human, obj>thing) and book(icl>thing).
  Here, the sense of the headword is focused by the attributes.
- ikebana(icl>flower_arrangement).
  Here, the headword comes from Japanese.
- go_down.
  Here, the headword does not need any refinement.

Ideally, an UW refers unambiguously to a concept, shared among several languages. However, UW are designed to represent acceptions in a language ; we therefore find distinct UW that refer to the same concept as for “affection” and “disease”.

We are mainly using the 207k UW built by the U++ Consortium (Jesus Cardeñosa et al., 2009) from the synsets of the Princeton WordNet, that are linked to natural languages via bilingual dictionaries. The storage of these dictionaries can be supported by a suitable platform like PIVAX (Nguyen et al., 2007) or a dedicated database. The gain of a pivot language is illustrated in figure 2. If we want to add a new language in the multilingual system, we just need to create the links with the pivot but not with all the other languages.

3.1.2 The Q-Systems

We can think of inserting the UW annotations with tags (e.g. XML) directly along the source text as in table 1. However, this naive approach is not adequate to represent the segmentation ambiguities that can occur in the text interpretation (in the example of table 1, we list the different possible meanings for “in”, but cannot represent “waiting”, “room” and “waiting room” as three possible lexical units).

In order to allow the representation of segmentation and other ambiguities, that can occur in a text interpretation, we propose to use the Q-Systems. They represent texts in an adequate

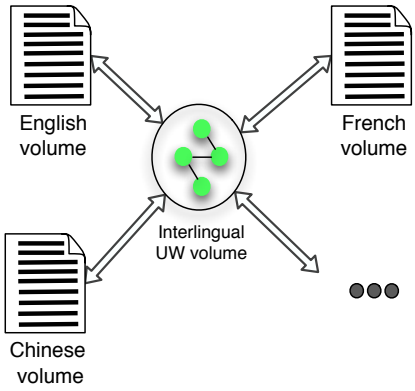

Figure 2: Multilingual architecture with a pivot

in a waiting room
<tag uw='in(icl-sup-how), in(icl-sup-adj), in(icl-sup-linear_unit, equu-sup-inch)'>**in**</tag> <tag uw='unk'>**a**</tag> <tag uw='waiting_room(icl-sup-room, equ-sup-lounge)'>**waiting room**</tag>

Table 1: Naive annotation of a text fragment

graph structure decorated with bracketed expressions (trees) and, moreover, allow processing on this structure via graph rewriting rules (a set of such rewriting rules is a so called Q-System).

An example of the Q-System formalism is given in figure 3 of section 3.2.3. It presents successively : the textual input representing a Q-graph, a rewriting rule and a graphical view of the Q-graph obtained after the application of the rule (and others).

The Q-Systems were proposed by Alain Colmerauer at Montreal University (Colmerauer, 1970). For our goal, they have three main advantages :

- they provide the formalized internal structure for linguistic portability that we mentioned in the introduction (Hajlaoui and Boitet, 2007) ;
- they unify text processing with powerful graph rewriting systems ;

- they allow the creation or the edition of a process by non-programmers (e.g. linguists) using SLLP (Specialized Language for Linguistic Programming).

We are actually using a reimplementation of the Q-Systems made in 2007 by Hong-Thai Nguyen during his PhD in the LIG-GETALP team (Nguyen, 2009).

3.2 Framework of the annotation process

3.2.1 Overview

The annotation process is composed by the following steps :

1. splitting the text in fragments if too long ;
2. lemmatisation with a specialized software ;
3. transcription to the Q-Systems format ;
4. creation of local bilingual dictionaries (source language - UW) for each fragment with PIVAX ;
5. execution of those dictionaries on the fragments ;

3.2.2 Lemmatisation

As we want to use dictionaries where entries are lemmas, the first step is to lemmatise the input text (i.e. to annotate occurrences with possible lemmas). This step is very important because it although gives the possible segmentations of the text in lexical units. It brings two kinds of ambiguities into play : on one hand, an occurrence can be interpreted as different lemmas, on the other, there can be several possible segmentations (eventually overlapping) to determine the lexical units.

For content extraction or information retrieval purpose, it is better to preserve an ambiguity than to badly resolve it. Therefore we expect from a lemmatiser to keep all ambiguities and to represent them in a confusion network (a simple tagger is not suitable). Several lemmatiser can be used to cover different languages. For each of them, we propose to use a dedicated ANTLR grammar (Terence Parr et al., 2009) in order to soundly transform the output in a Q-graph.

To process the Belga corpus, we developed a lemmatiser that produce natively Q-graphs. It is based on the morphologic dictionary DELA[1] available under LGPL licence.

3.2.3 Local dictionaries as Q-Systems

Having the input text annotated with lemmas, with the Q-System formalism, we want to use the graph rewriting possibilities to annotate it with UW. To do so, we use PIVAX export features to produce rules that rewrite a lemma in an UW (see figure 3). Each rule correspond to an entry in the bilingual dictionary. To obtain a tractable Q-Systems (sets of rules), we built local dictionaries that contain the entries for fragments of the text (about 250 words in the first experiment).

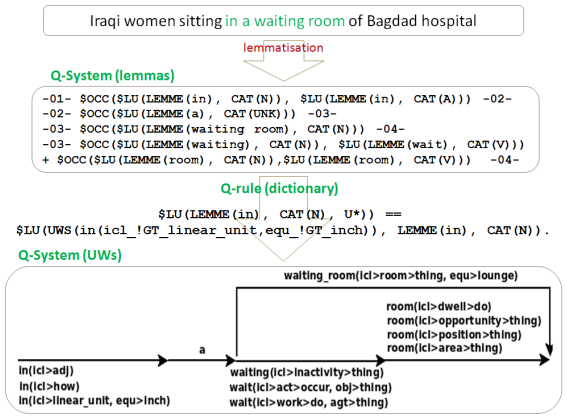

Figure 3: Creation and execution of a Q-System

Considering the significant quantity of ambiguities generated by this approach (up to a dozen UW for a single word), we need to include a disambiguation process. This process, based on conceptual vectors, is presented in the next section.

4 *Conceptual vector based disambiguation*

Vectors have been used in NLP for over 40 years. For information retrieval, the standard vector model (SVM) was invented by Salton (Salton, 1991) during the late 60's, while for meaning representation, latent semantic analysis (LSA)

[1] http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html

was developed during the late 80's (Deerwester et al., 1990). These approaches are inspired by distributional semantics (Harris et al., 1989) which hypothesises that a word meaning can be defined by its co-text. For example, the meaning of ‹*milk*› could be described by {‹*cow*›, ‹*cat*›, ‹*white*›, ‹*cheese*›, ‹*mammal*›, ...}. Hence, distributional vector elements correspond directly (for SVM) or indirectly (for LSA) to lexical items from utterances.

The conceptual vector model is different as it is inspired by componential linguistics (Hjelmlev, 1968) which holds that the meaning of words can be described with semantic components. These can be considered as atoms of meaning (known as primitives (Wierzbicka, 1996)), or also only as constituents of the meaning (known as semes, features (Greimas, 1984), concepts, ideas). For example, the meaning of ‹*milk*› could be described by {*LIQUID*, *DAIRY PRODUCT*, *WHITE*, *FOOD*, ...}. Conceptual vectors model a formalism for the projection of this notion in a vectorial space. Hence, conceptual vector elements correspond to concepts indirectly, as we will see later.

For textual purposes[2], conceptual vectors can be associated to all levels of a text (word, phrase, sentence, paragraph, whole texts, etc.). As they represent ideas, they correspond to the notion of *semantic field*[3] at the lexical level, and to the overall thematic aspects at the level of the entire text.

Conceptual vectors can also be applied to lexical meanings. They have been studied in word sense disambiguation (WSD) using isotopic properties in a text, i.e. redundancy of ideas (Greimas, 1984). The basic idea is to maximise the overlap of shared ideas between senses of lexical items. This can be done by computing the angular distance between two conceptual vectors (Schwab and Lafourcade, 2007).

In our case, conceptual vectors are used for automatic disambiguation of texts. Using this method, we calculate confidence score for each UW hypothesis appearing in the Q-Graph.

[2]Conceptual vectors can be associated with any content, not only text: images, videos, multimedia, Web pages, etc.

[3]The semantic field is the set of ideas conveyed by a term.

5 *Ontology driven content extraction*

The content extraction has to be leaded by a "knowledge base" containing the informations we want to retrieve.

5.1 Previous works in content extraction

This approach has its roots in machine translation projects such as C-Star II (1993-1999) (Blanchon and Boitet, 2000) and Nespole! (2000-2002) (Metze et al., 2002), for on the fly translation of oral speech acts in the domain of tourism. In these projects, semantic transfer was achieved through an IF (Inter-exchange Format), that is a semantic pivot dedicated to the domain. This IF allows to store information extracted from texts but is although used to lead the content extraction process by giving a formal representation of the relevant informations to extract, according to the domain.

The Nespole! IF consists of 123 concepts from the tourism domain, associated with several arguments and associable with speech acts markers. The extraction process is based on patterns. As an example, the statement "*I wish a single room from September 10th to 15th*" may be represented as follows:

```
{ c:give-information+disposition+room
  ( disposition=(desire, who=i),
    room-spec=
    ( identifiability=no,single_room ),
    time=
    ( start-time=(md=10),
      end-time(md=15, month=9)
    )
  )
}
```

5.2 Ontologies as parameter for the domain

In the project OMNIA, the knowledge base has the form of a lightweight ontology for image classification [4]. This ontology contains 732 concepts in the following domains : animals, politics, religion, army, sports, monuments, transports, games, entertainment, emotions, etc. To us, using an ontology has the following advantages :

- Ontologies give an axiomatic description of a domain, based on formal logics (usu-

[4]http://kaiko.getalp.org/kaiko/ontology/OMNIA/OMNIA_current.owl

ally description logics (Baader et al., 2003)) with an explicit semantic. Thus, the knowledge stored in them can be used soundly by software agents;

- Ontological structures are close to the organisation of ideas as semantic networks in human mind (Aitchenson, 2003) and are labeled with strings derived from natural languages. Thus humans can use them (browsing or contributing) in a pretty natural way;

- Finally, with the advent of the Semantic Web and normative initiatives such as the W3C[5], ontologies come with a lot of shared tools for editing, querying, merging, etc.

As the content extractor might only process UW annotations, it is necessary that the knowledge base is whether expressed using UW or linked to UW. The ontology is here considered as a domain parameter of content extraction and can be changed to improve preformances on specific data collections. Therefore, given any OWL ontology[6], we must be able to link it with a volume of UW considering the following constraints :
Creating manually such correspondences is costly due to the size of resources so an automatic process is requiered.
Ontologies and lexicons evolve over the time so an alignment must be adaptable to incremental evolutions of resources.
The correspondences must be easily manipulated by users so they can manually improve the quality of automatically created alignments with post-edition.

Constructing and maintaining an alignment between an ontology and an UW lexicon is a challenging task (Rouquet and Nguyen, 2009b). Basically, any lexical resource can be represented in an ontology language as a graph. We propose to use an OWL version of the UW volume available on Kaiko website [7]. It allows us to benefit of classical ontology matching techniques and tools (Euzenat and Shvaiko, 2007) to represent, compute and manipulate the alignment. We implemented two string based matching techniques on top of the alignment API (Euzenat, 2004). Specific disambiguation methods are in development to improve the alignment precision. Some of them are based on conceptual vectors presented in section 4, others will adapt structural ontology matching techniques. This approach to match an ontology with a lexical resource is detailled in (Rouquet et al., 2010).

5.3 The generic extractor

In the case of the OMNIA project, the system output format is constraint by the goal of an integration with visual analysis results, in a larger multimodal system. The visual analysis systems are also based on concept extraction, but does not need an ontology to organise concepts. Therefore, our results has to remain autonaumous, which means without references to the ontology used to extract concepts. So, we use a simple concept vector as output, with intensity weights; practically, a simple data-value pairs sequence formatted in XML.

Concept extraction is achieved through a 3 steps process, has shown in figure 4.

1. *Concept matching*: each UW in the Q-Graph, that matches a concept according to the UW-concept map, is labelled with this concept.

2. *Confidence calculation*: each concept label is given a confidence score, in accordance with the score of the UW carrying the concept, obtained after disambiguation, and pondered according to the number of UWs in the Q-Graph. It is planed to take into account a few linguistics hints here, such as negations, and intensity adverbs.

3. *Score propagation*: because we need autonomous results, we have to perform all ontology-based calculation before releasing them. The confidence scores are propagated in the ontology concept hierarchy: for each

[5] http://www.w3.org/
[6] http://www.w3.org/2004/OWL/
[7] http://kaiko.getalp.org

labelled concept, its score is added to the super-concept, and so on.

The ontology and the derivated UW-concept map are considered as parameters for the treatments, and may be replaced in accordance with the domain, and the relevance of the concepts and their hierarchy, according to the task.

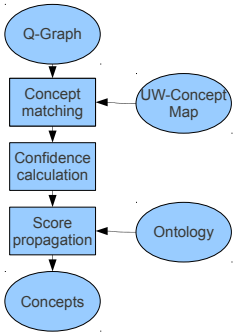

Figure 4: Detail of concept extraction.

6 Experiments

For a first experiment, we used a small dataset, containing:

- a sub-corpus of 1046 English companion texts from CLEF09 corpus (press pictures and captions of about 50 words),
- a 159 concepts ontology, designed for picture and emotions depiction,
- a UW-concept map comprising 3099 UW.

It appeared that, with this parameters, concepts where extracted for only 25% of the texts. This preliminary result stressed the importance of recall for such short texts. However, there were many ways to improve recall in the system:

- improve the ontology, in order to better cover the press domain;
- significantly increase the quantity of UW linked to concepts (only 3099 obtained for this experiment), by considering synonyms during the linking process;
- using UW restrictions during concept matching for UW that are not directly linked to a concept, as these restrictions are a rich source of refined semantic information.

A second experiment with an improved ontology, including 732 concepts, and the use of UW restrictions, showed very promising results. Concepts were retrieved from 77% of texts. The remaining texts were very short (less than 10 words, sometime just date or name).

For example, we extracted the following concepts from the picture and companion text reproduced in figure 5.

AWA05 - 20020924 - BAGHDAD, IRAQ : Iraqi women sit under a portrait of Iraqi President Saddam Hussein in a waiting room in Baghdad's al-Mansur hospital 24 September 2002. Saddam Hussein is doggedly pursuing the development of weapons of mass destruction and will do his best to hide them from UN inspectors, the British government claimed in a 55-page dossier made public just hours before a special House of Commons debate on Iraq. Iraqi Culture Minister Hamad Yussef Hammadi called the British allegations "baseless." EPA PHOTO AFPI AWAD AWAD

Figure 5: Picture document and companion text example.

CONCEPT	WEIGHT
BUILDING	0.098
HOSPITAL	0.005
HOUSE	0.043
MINISTER	0.016
OTHER_BUILDING	0.005
PEOPLE	0.142
PERSON	0.038
POLITICS	0.032
PRESIDENT	0.016
RESIDENTIAL_BUILDING	0.043
WOMAN	0.005

As this results were more consistent, we could have a preliminary survey about precision, on a 30 texts sample. While disambiguation implementation is still at an early stage, weights were not yet taken into account. A concept match can be considered correct following two criterons :

1. **Visual relevance** considers a concept as correct if carried by an element of the picture; for instance, the match of concept

"SPORT" is regarded as correct for a picture containing a minister of sports, even if not actually performing any sport.

2. **Textual relevance** considers a concept as correct if carried by a word of the text, as parts of texts may involve concepts that are not actually present in the picture, such as contextual information, previous events, etc.

124 concepts were found in 23 texts (7 texts had no concept match):

1. 99 concepts were correct according to the visual relevance,
2. 110 were correct according to the textual relevance,
3. 14 were totally incorrect.

We thus have an overall precision score of 0.798 according to the visual relevance and 0.895 according to the textual relevance. Most of the errors where caused by ambiguity problems, and may be addressed with disambiguation process that are not fully implemented yet.

7 Conclusion and perspectives

We exposed a generic system designed to extract content (in the form of concepts) from multilingual texts. Our content extraction process is generic regarding to two aspects :

- it is language independent, as it process an interlingual representation of the texts
- the content to be extracted can be specified using a domain ontology as a parameter

This is an ongoing work, and disambiguation through conceptual vectors is expected to improve accuracy, giving significant weights to the hypothetical meanings of words.

In the long run, we will focus on integration with visual content extractors, speed optimization to achieve a real-time demonstrator and detailled evaluation of the method.

References

Aitchenson, J. 2003. *Words in the Mind. An Introduction to the Mental Lexicon*. Blackwell Publishers.

Baader, De Franz, Diego Calvanese, Deborah McGuinness, Peter Patel-Schneider, and Daniele Nardi. 2003. *The Description Logic Handbook*. Cambridge University Press.

Blanchon, H. and C. Boitet. 2000. Speech translation for french within the C-STAR II consortium and future perspectives. In *Proc. ICSLP 2000*, pages 412–417, Beijing, China.

Boitet, Christian, Igor Boguslavskij, and Jesus Cardeñosa. 2009. An evaluation of UNL usability for high quality multilingualization and projections for a future UNL++ language. In *Computational Linguistics and Intelligent Text Processing*, pages 361–373.

Colmerauer, A. 1970. Les systèmes-q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. *département d'informatique de l'Université de Montréal, publication interne*, 43, September.

Daoud, Daoud. 2006. *Il faut et on peut construire des systèmes de commerce électronique à interface en langue naturelle restreints (et multilingues) en utilisant des méthodes orientées vers les sous-langages et le contenu*. Ph.D. thesis, UJF, September.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6).

Euzenat, Jérôme and Pavel Shvaiko. 2007. *Ontology matching*. Springer, Heidelberg (DE).

Euzenat, Jérôme. 2004. An API for ontology alignment. In *Proceedings of the 3rd International Semantic Web Conference*, pages 698–7112, Hiroshima, Japan.

Fielding, Roy T. 2000. *Architectural styles and the design of network-based software architectures*. Ph.D. thesis, University of California.

Greimas, Algirdas Julien. 1984. *Structural Semantics: An Attempt at a Method*. University of Nebraska Press.

Hajlaoui, Najeh and Christian Boitet. 2007. Portage linguistique d'applications de gestion de contenu. In *TOTh07*, Annecy.

Harris, Zellig S., Michael Gottfried, Thomas Ryckman, Paul Mattick Jr., Anne Daladier, T.N. Harris, and S. Harris. 1989. *The form of Information in Science, Analysis of Immunology Sublanguage*, volume 104 of *Boston Studies in the Philosophy of Science*. Kluwer Academic Publisher, Dordrecht.

Hjelmlev, Louis. 1968. *Prolégolème à une théorie du langage*. éditions de minuit.

Jesus Cardeñosa et al. 2009. The U++ consortium (accessed on september 2009). http://www.unl.fi.upm.es/consorcio/index.php, September.

Luca Marchesotti et al. 2010. The Omnia project (accessed on may 2010). http://www.omnia-project.org, May.

Max Silberztein. 2009. NooJ linguistic software (accessed on september 2009). http://www.nooj4nlp.net/pages/nooj.html, September.

Metze, F., J. McDonough, H. Soltau, A. Waibel, A. Lavie, S. Burger, C. Langley, L. Levin, T. Schultz, F. Pianesi, R. Cattoni, G. Lazzari, N. Mana, and E. Pianta. 2002. The Nespole! speech-to-speech translation system. In *Proceedings of HLT-2002 Human Language Technology Conference*, San Diego, USA, march.

Nguyen, H.T., C. Boitet, and G. Sérasset. 2007. PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot. In *SNLP*, Bangkok, Thailand.

Nguyen, Hong-Thai. 2009. EMEU_w, a simple interface to test the Q-Systems (accessed on september 2009). http://sway.imag.fr/unldeco/SystemsQ.po?localhost=/home/nguyenht/SYS-Q/MONITEUR/, September.

Rouquet, David and Hong-Thai Nguyen. 2009a. Interlingual annotation of texts in the OMNIA project. Poznan, Poland.

Rouquet, David and Hong-Thai Nguyen. 2009b. Multilinguïsation d'une ontologie par des correspondances avec un lexique pivot. In *TOTh09*, Annecy, France, May.

Rouquet, David, Cassia Trojahn, Didier Scwab, and Gilles Sérasset. 2010. Building correspondences between ontologies and lexical resources. In *to be published*.

Salton, Gerard. 1991. The Smart document retrieval project. In *Proc. of the 14th Annual Int'l ACM/SIGIR Conf. on Research and Development in Information Retrieval*, Chicago.

Schwab, Didier and Mathieu Lafourcade. 2007. Lexical functions for ants based semantic analysis. In *ICAI'07- The 2007 International Conference on Artificial Intelligence*, Las Vegas, Nevada, USA, juin.

Terence Parr et al. 2009. ANTLR parser generator (accessed on september 2009). http://www.antlr.org/, September.

Uchida Hiroshi et al. 2009. The UNDL foundation (accessed on september 2009). http://www.undl.org/, September.

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals*. Oxford University Press.