Multilingual Lexicons for Machine Translation

LIM Lian Tze
liantze@gmail.com
Faculty of Information Technology
Multimedia University, Cyberjaya, Selangor, Malaysia

ABSTRACT

Despite the numerous bad press received by machine translation systems, they are invaluable in aiding users to quickly gather the essence of texts written in an unfamiliar language, particularly on the Web. Sufficiently accurate translation of lexical items is essential to satisfy such needs. Lexicon resources, especially a well-designed multilingual lexicon is needed to facilitate effective translation selection of lexical items, consisting of either single or multiple words. We describe the issues facing multilingual lexicography, and review how they are handled in selected multilingual lexical database projects. This preliminary surveys aims at identifying multilingual issues and concerns in translating lexical items from a machine translation perspective, and will guide the design of a multilingual lexicon intended for use in a machine translation system.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Dictionaries, linguistic processing; I.2.7 [**Natural Language Processing**]: Machine translation

General Terms

Design

Keywords

Multilingual lexical databases, machine translation

1. INTRODUCTION

Machine translation (MT) systems are computer programs that automatically translate natural language text from a source language (SL) to a target language (TL). MT is difficult not only because each language differs from the next (even those from the same family) both structurally and lexically, but also because natural language is itself ambiguous (again, both structurally and lexically) and always evolving. As such, MT has received much bad press due to unrealistic public expectations that MT systems should produce publishable-quality, no-further-improvements-required translations at the press of a button.

1.1 Usage Context of Machine Translation

The real value of MT technology is apparent when its usage context is viewed correctly. Hovy [6] and Hutchins [8] identified three usage scenarios of MT where human end-users are concerned:

Dissemination Producing a translation 'draft' to be manually post-edited to publishable quality.

Assimilation 'Gisting', or aiding users to find out essential contents of a document. Lower quality is expected and acceptable.

Interchange/Communication Immediate translation to convey basic contents of messages in multi-turn dialogue, such as telephone conversations and chats.

Hutchins further listed **information access** as a usage context, where MT is integrated into other computer systems, like cross-lingual query and retrieval systems.

Using MT for assimilation and information access purposes is particularly relevant in this "information age". The amount of information available on the Web has grown to an unimaginable size, at an equally astounding speed. More often than not, however, these information are not available in languages an end-user is comfortable with. This is especially true for user-generated content such as blog posts and discussion threads, which are unlikely to be manually translated into other languages by the original author (or any other party).

1.2 Translation Selection

A translation may satisfy assimilation needs if it contains fairly accurate lexical items, even if the output is not syntactically well-formed. Take, for example, the following Welsh input text and its translation output by an online Welsh–English MT system at `http://www.cymraeg.org.uk`[1]:

Input Cafodd gyrrwr a fethodd brawf anadl cyn ymosod ar blismon a gyrru i ffwrdd ar gyflymder o 100 m.y.a. ei garcharu am 27 mis.

Output Driver got and failed *brawf breath before attack on *blismon and drive to a way on a speed of 100 *m.the.and. imprison him for 27 months.

iiWAS2009, December 14–16, 2009, Kuala Lumpur, Malaysia.

[1] This example was given by Mikel L Forcada in a comment posted at `http://blogs.ft.com/brusselsblog/2009/01/cheeseburgery-hamburgers-and-the-problem-of-computerised-translations/`

The same is true for information access purposes, particularly cross-lingual search and retrieval. A user who specifies search keywords in language L_1 would be able to get results in other languages if the keywords are translated via an embedded MT module. The cross-lingual results would only be relevant if the keywords are translated correctly.

For accurate *translation selection*, that is, selection of TL lexical items to build the translation output, MT systems require well-structured lexical resources with sufficiently broad coverage. In the following sections, we first describe the types of lexical resources useful for MT. We then mention some issues related to multilingualism, and review how they are handled in some multilingual lexical database projects.

1.3 Scope of Study

This is a preliminary survey, undertaken as a first step to study the implications of multilingualism in computational lexicography and MT. The insights gained from this preliminary survey will guide the design of a computer-tractable multilingual lexicon for use in an MT system. Following is the overall proposed research plan:

- Study the literature and data to identify multilingual issues.
- Design a data structure for describing multilingual translation equivalence, based on the findings.
- Develop a proof-of-concept multilingual lexicon from bilingual dictionaries:
  - pre-process bilingual dictionaries to extract fields,
  - align sense entries from bilingual sources to pivots,
  - annotating multi-level correspondences on multi-word translation equivalents,
  - generating semantic information for translation selection purposes.
- Evaluating the multilingual lexicon in a MT-for-assimilation context, against an existing MT system:
  - develop a translation selection module to be embedded in an existing MT system, based on the data structures (including supporting semantic information) in the multilingual lexicon and user's editing actions,
  - fidelity tests (users rate the adequacy of translations),
  - comprehensibility tests (users answers multiple-choice questionnaires based on translations).

2. LEXICAL RESOURCES FOR MT

We briefly describe the characteristics of lexical resources that would be useful for MT systems. We are interested in resources providing information about the general lexicon of languages, rather than discipline-specific terminologies.

2.1 Monolingual Resources

Monolingual lexical resources are required in MT systems to analyse the SL input and to generate well-formed output in the TL. They provide information on the syntactic and morphological behaviors of lexical items in a specific language, such as part-of-speech (POS), sub-categorisation frame, case, gender, number and tense. With these information, parser modules in MT systems can identify the lemma or canonical forms of individual lexical items in the input text, select lexical items (in their lemmatised forms) in the TL (see following section), and produce morphologically correct forms to build the final translation.

The monolingual lexical resource should also take into account alternate surface forms, especially if a language has multiple writing systems. For example, the Japanese lexical item for 'world' can be written as kanji '世界', hiragana 'せかい' or katakana 'セカイ' (all transliterated as *sekai*); or that '*organize*' and '*organise*' are American and British spellings respectively of the same word. The monolingual lexical resource should thus provide all alternate surface forms possible for a lexical item, so that all alternate forms would be treated equally when the input text is parsed. In addition, it should also list multi-word expressions (MWEs) in the language, including non-contiguous ones.

How language-specific morphological and syntactic behaviors should be modelled and stored is outside the scope of this paper; see [10] instead for a comprehensive account.

As mentioned earlier, natural languages contain lexical ambiguities. For instance, the English noun '*bank*' may mean a financial institution or the land beside a body of water, and thus would be translated differently depending on the context. A monolingual lexical resource containing semantic information would be helpful for an MT system to disambiguate polysemous lexical items in the input text. Note that such semantic information for disambiguation purposes may also be associated to entries in a bilingual or multilingual resource instead: this may be more desirable and relevant in MT systems [5, 13, 16].

The actual type of semantic information used may vary depending on the disambiguation approach used, as well as the paradigm of the MT system itself. There is a rich literature on word sense disambiguation (WSD) and its supporting resources, including corpus statistics and probabilities, selectional preferences, taxonomic relations between lexical senses, topic signatures, domain or subject field codes, etc. See [15] for a more detailed overview.

Gloss texts of word senses may also be used to generate data for disambiguation purposes [1, 12, 14] although not compulsory. On the other hand, they can be incorporated as an extra dictionary look-up option for human users if the MT system is part of a translator workbench.

2.2 Bilingual and Multilingual Resources

MT systems must be equipped with repositories of lexical translational equivalents across languages. The simplest scheme is a "flat" list mapping L_1 lexical items to one or more possible translations in the target language L_2, sometimes not even making any distinctions between different senses (Figure 1). Such lists are actually unidirectional, thus an MT system that translates in both directions between a language pair would require two such translation lexicons.

POS	English	Malay
n	bank	bank; tabung; tebing; beting; tambak; permatang
v	bank	menyimpan wang; menimbun; terbang mengereng

Figure 1: Simple English–Malay bilingual translation lexicon without sense distinctions

While this bilingual scheme is easy to maintain for a MT system that handles a single language pair (requiring two unidirectional bilingual lexicons), the number of inter-lexicon links to maintain grows quickly to $O(n^2)$ in a system involving n languages, as shown in Figure 2. Adding a new language requires $O(n)$ new links to link the new language to each of the already existing languages.

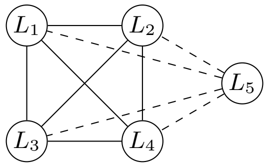

Figure 2: Adding a new language in bilingual lexicons setting

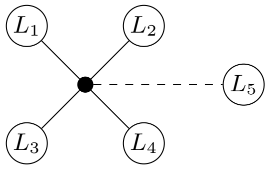

Figure 3: Adding a new language in multilingual lexicon setting

On the other hand, a system using a *multilingual* lexicon requires the maintenance of only $O(n)$ links to a *pivot* (Figure 3). Entries for a new language would only need to be linked to the pivot, and translation equivalence between the new language and the existing ones would be established via the pivot. This is especially beneficial to introduce more language pairs, especially from and to less-resourced languages, into a MT system. At a glance, the pivot is similar to an interlingua. Despite the various objections to the existence of a universal language (mainly from researchers in linguistics and psychology [3, 7]), such a mechanism presents a feasible solution if treated as a computational mechanism rather than for explaining fundamental linguistic issues [18].

Due to linguistic phenomena and differences across languages, however, merging bilingual lexicons into a single multilingual repository is non-trivial. A lexical resource that aims at providing multilingual translation equivalents must be well-designed to address these issues. We will briefly mention some of the related problems in the next section.

3. ISSUES IN MULTILINGUAL LEXICOGRAPHY

Given the abundant existence of bilingual lexicons, creating a multilingual lexicon, although seemingly straightforward, can be fraught with inherent linguistic problems. Hutchins [9] named two main issues related to bilingual lexical differences. The first issue concerns bilingual lexical ambiguity, or the existence of multiple equivalents in the target language. This could be due to ambiguity in the source language, for example English '*glass*'→ '*gelas*' (a receptacle for fluids) and '*glass*'→'*kaca*' (a material made from sand) in Malay. In other cases, a single meaning of a lexical item may have translations in the target language that are more specific, such as the Spanish '*dedo*'→'*finger*' and '*dedo*'→'*toe*' in English as it does not distinguish between appendages on the hand or foot. This phenomena is known as diversification, and as neutrification in the opposite direction.

The second issue mentioned by Hutchins is that of lexical gaps, when a concept is not lexicalised in a particular language. This is sometimes due to cultural differences: indeed many words pertaining to culinary or clothing apparels in a specific culture do not have equivalents in other languages, like '*cottage*', '*vodka*', '*batik*', '粽子' (*zòngzi* Chinese glutinous rice dumpling), 'きもの' (*kimono*). Romanised or transliterated forms of such lexical items are usually used in the translated text, and often find their way into the TL's vocabulary. In most other cases, a gloss-like expression or a paraphrase is used to translate the SL lexical item. For example, the English adjective '*absent*' is translated as an adjectival phrase '*tidak hadir*' (not present) in Malay. However, it would be uneconomical to store a gloss for every lexical gap, especially if a concept is specific to a particular (family of) culture or language.

Occasionally an SL lexical item and one from the TL may be very near synonyms, yet have subtle underlying differences, resulting in near-miss lexical gaps in both languages. Consider Chinese '跳飞机' (*tiào fēijī*) and Indonesian '*merantau*': while both describe a situation where a person works in a foreign country without intentions to reside permanently, the former has a negative connotation while the latter does not. Such subtle differences often confuse the TL-speaking user and annoys the SL-speaker. This is perhaps unavoidable, as a human professional translator may have no better strategy but to offer the same translation.

A third issue is when either of the mapped expressions (both SL and TL) contain multiple words,[2] they may not necessarily be contiguous. In a conventional bilingual dictionary, the translation equivalents are described using human-recognisable place-holders, for example:

- English: '*throw* ***somebody*** *to the lions*'
  Chinese: '丢下某人不管'
- English: '*pull* ***one's*** *weight*'
  Malay: '*turut bekerja keras*'
- English: '*get* ***one's*** *knife into* ***somebody***'
  Malay: '*berniat jahat terhadap* ***seseorang***'

Such linear sequences may be inadequate in a multilingual setting, where the correspondences between substrings of decomposable translation equivalents in different languages are more complex.

4. MULTILINGUAL LEXICON DESIGNS

The idea of using a particular natural language as a linking pivot may seem feasible at first glance: after all, trilingual dictionaries *do* exist: the FeM dictionary (French–English–Malay) at `http://www-clips.imag.fr/geta/services/fem` is one. However, such an approach is suitable only for human consultation and poses problems to computational processing.

English	French	Malay	Japanese
rice	riz	padi	稲
rice	riz	beras; nasi	米；御飯

Figure 4: Problem with using a natural language as a linking pivot

Figure 4 shows sample entries from a multilingual lexicon using English as the pivot. Two senses of '*rice*', namely the plant itself and the grains of the plant, are listed together with their translations in French, Malay and Japanese. However, Malay and Japanese distinguish between *cooked* and *uncooked* rice grains, using '*nasi*' and '御飯' (*gohan*) for the former, '*beras*' and '米' (*yone*) for the latter. This is due to the diversification phenomena mentioned in section 3. Should this multilingual lexicon be used in an MT system, either '御飯' or '米' would be equally likely to be selected to translate '*beras*'. In addition, if a concept is not lexicalised in English (or whatever the pivot language is), lexical items in the other languages cannot be entered into the lexicon at all. Therefore,

[2] the mapped TL string may either be a lexical item (MWE) or not.

any design that uses an actual natural language as the linking pivot is not feasible for multilingual MT purposes.

4.1 EuroWordNet

EuroWordNet [17] uses a language-independent Inter-Lingual Index (ILI) to link synonymous lexical senses in different languages, using English for convenient naming of the ILI records. Recall the earlier example on Spanish '*dedo*' (and also Italian '*dito*') having more specific translations in English '*toe*' and '*finger*'. English '*toe*' and '*finger*' are linked to the respective ILI records using normal equivalence relations, while '*dito*' and '*dedo*' are linked as equivalence to a separate ILI record. '*dito*' and '*dedo*' are further linked to **toe** and **finger** ILI records using hyponym-equivalence (more general than) relations. Note that the ILI records are not structured in any way. Such use of hyponym-equivalence and respectively hypernym-equivalence (more specific than) can handle diversification and neutrification. However, as shown in Figure 5, EuroWordNet's ILI design would cause an explosion of links to maintain when records similar to **dedo, dito** are created. In addition, we are not aware of any provisions for non-contiguous MWEs in EuroWordNet.

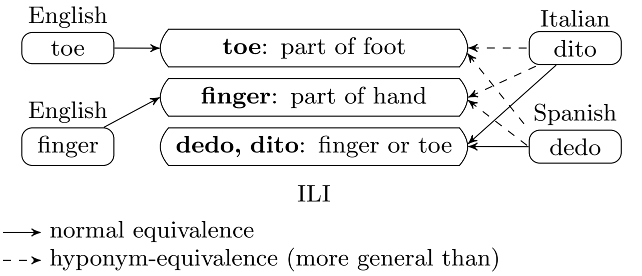

Figure 5: EuroWordNet's Unstructured Inter-Lingual Index (after Vossen [17])

4.2 Papillon

The Papillon multilingual dictionary project [2] uses a volume of interlingual *axies* to link translation equivalents from different languages. As Papillon's axies may have relations among themselves, contrary to EuroWordNet's ILI records, the problem of 'link explosion' can be avoided. This is illustrated in Figure 6, where the 'grain' sense of '*rice*' and '*riz*' are linked to an axie that is further linked to two other axies. '米' and '*beras*' (respectively '御飯' and '*nasi*') can then be specified as equivalent to each other, and are more specific than '*rice*' and '*riz*'. We are also unaware of any provisions for non-contiguous MWEs in Papillon.

4.3 Lexical Markup Framework

The Lexical Markup Framework (LMF) [4, 10] was introduced as an ISO standard for lexical resource management (ISO 24613) and provides mechanisms for various aspects of lexicography related to natural language processing, including morphology, syntax, semantics and multilingualism. The Sense Axis in LMF is similar in nature to Papillon's axie.[3] LMF also has a Transfer Axis for specifying multilingual translation equivalents with selectional restriction tests. For example, English '*to develop*' is translated to Italian '*construire*' and Spanish '*construir*' if the second syntactic argument is a building; otherwise it is translated to the more general Spanish '*desarrollar*'. On the other hand, there are mechanisms in LMF for specifying MWEs and their possibly non-contiguous and decomposable constructions, but how the correspondences between the components of their translations are handled are not stated.

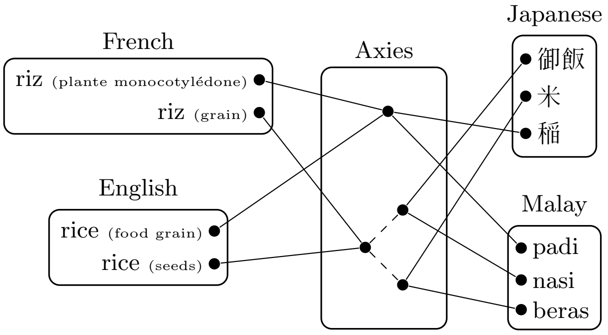

Figure 6: Papillon's interlingual axies (after Boitet, Mangeot, and Sérasset [2])

4.4 SIM*u*LLDA

In the multilingual lexicon projects and frameworks reviewed so far, the interlingua is a pivot structure for linking synonymous senses of lexical items from different languages. As such, when there is a lexical gap in a particular language L, a translation can only be generated for L by translating the *gloss text*, if available, of a lexical item in another language. SIM*u*LLDA [11] takes a different approach by using a taxonomic lattice of concepts or definitional attributes as the interlingua, based on Formal Concept Analysis principles. MWEs were not considered in [11].

In the example on lexical items related to *horse* in Figure 7, there is a lexical gap in French for English '*colt*'. From the lattice of concepts and definitional attributes, '*colt*' ≡ COLT = FOAL + **male**. There *is* a French equivalent for **foal**: '*poulain*'. A French translation can therefore be systematically generated, i.e. '*poulain mâle*'.

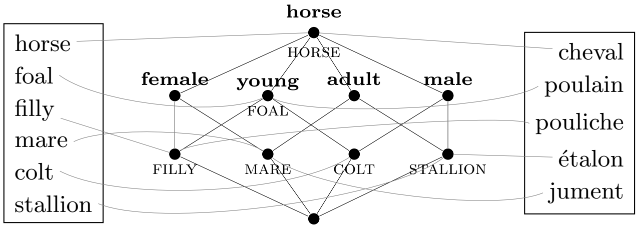

Figure 7: SIM*u*LLDA's lattice of concepts and definitional attributes (after Janssen [11])

However, SIM*u*LLDA's taxonomic considerations do not always agree with lexicographic practices. Translation equivalence cannot be established among many accepted translation pairs if strict logical principles are applied, or would be problematic if it is attempted: see Janssen's elaboration on French '*rivière*', '*fleuve*' and English '*river*', '*stream*' in [11]. Translational equivalence is not based on logical principles; it is a phenomenon borne of acceptance and common practice

[3] LMF borrowed the term 'axie' from Papillon but changed it to 'axis' to respect English orthography [4].

by human speakers of natural languages. The best role of multilingual lexicons for MT systems is to capture this notion of equivalence as perceived by human speakers, rather than to apply strict formal analysis or operations on lexical senses.

5. CONCLUSION

The translation selection module in an MT system must accurately generate lexical translation equivalents, which will go far in satisfying assimilation and information access needs. The system thus needs well-designed multilingual lexicons capable of addressing multilingual issues.

We described three issues in multilingual lexicography: multiple translation equivalents in the TL due to SL ambiguity and diversification; lexical gaps; and non-contiguous, decomposable MWEs. We also reviewed how they are addressed in a selection of multilingual lexicon projects and frameworks, namely EuroWordNet, Papillon, LMF and SIM*u*LLDA.

We conclude that a language-independent pivot structure that supports diversification links, such as those in Papillon or LMF, is the best approach for capturing translation equivalence. Formal or logical analysis methods, such as SIM*u*LLDA's approach, can be problematic as translation is not necessarily a "logical" operation. We also note that correspondences to translations of non-contiguous and decomposable MWEs, as well as multi-word forms that are not lexical items in the TL, do not receive much attention in the reviewed multilingual lexicons. We believe that this issue merits further study so that MT systems are more robust at translating MWEs and lexical gaps.

Based on these conclusions, a multilingual lexicon for MT will be designed with at least the following characteristics:

- a language-independent pivot structure with diversification features,
- capable of describing irregular correspondences in multi-word translation equivalents.

A translation module using a proof-of-concept lexicon will be added to an existing MT system, which will be evaluated in an MT-for-assimilation context.

6. ACKNOWLEDGEMENTS

The work described in this paper was supported by a ScienceFund grant (01-02-01-SF0117) from the Malaysian Ministry of Science and Technology. The author would like to thank Dr Tang Enya Kong and Dr Bali Ranaivo-Malançon for their supervision, the three anonymous reviewers for their comments, and Prof. Wolfram Wöß for shepherding this paper.

7. REFERENCES

[1] S. Banerjee and T. Pedersen. "Extended Gloss Overlaps as a Measure of Semantic Relatedness." In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence.* 2003, pp. 805–810.

[2] C. Boitet, M. Mangeot, and G. Sérasset. "The PAPILLON project: Cooperatively Building a Multilingual Lexical Database to Derive Open Source Dictionaries & Lexicons." In: *Proceedings of the 2nd Workshop on NLP and XML (NLPXML'02).* Taipei, Taiwan 2002, pp. 1–3.

[3] M. H. Christiansen and N. Chater. "Language as Shaped by the Brain." In: *Behavioral and Brain Sciences* 31.5 (2008), pp. 489–509.

[4] G. Francopoulo et al. "Multilingual Resources for NLP in the Lexical Markup Framework (LMF)." In: *Language Resources and Evaluation* 43.1 (Mar. 2009), pp. 57–70. DOI: 10.1007/s10579-008-9077-5.

[5] W. Gale, K. W. Church, and D. Yarowsky. "Using Bilingual Materials to Develop Word Sense Disambiguation Methods." In: *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation.* Montreal, Canada 1992, pp. 101–112.

[6] E. Hovy. "Toward Finely Differentiated Evaluation Metrics for Machine Translation." In: *Proceedings of the EAGLES Workshop on Standards and Evaluation.* Pisa, Italy 1999.

[7] J. R. Hurford. "Nativist and Functional Explanations in Language Acquisition." In: *Logical Issues in Language Acquisition.* Ed. by I. M. Roca. Dordrecht, the Netherlands: Foris Publications, 1990.

[8] J. Hutchins. "The Development and Use of Machine Translation Systems and Computer-based Translation Tools." In: *Proceedings of the International Symposium on Machine Translation and Computer Language Information Processing.* Beijing, China 1999, pp. 1–16.

[9] J. Hutchins. *Machine translation: problems and issues.* Presentation. Chelyabinsk, Russia. 18 slides. 2007.

[10] International Organization for Standardization. *ISO 24613:2008 Language Resource Management – Lexical Markup Framework (LMF).* 2008.

[11] M. Janssen. "Lexical Translation and Conceptual Hierarchies." In: *Proceedings of the 5th International Symposium on Language, Logic and Computation.* Tbilisi, Georgia 2003.

[12] M. Lesk. "Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone." In: *Proceedings of the ACM-SIGDOC Conference.* Toronto, Canada 1986, pp. 24–26.

[13] L. T. Lim. "Improving Translation Selection with Conceptual Vectors." In: *Proceedings of the Regional Computer Science Postgraduate Colloquium (ReCSPC'06).* Penang, Malaysia 2006, pp. 231–234.

[14] L. T. Lim and D. Schwab. "Limits of Lexical Semantic Relatedness with Ontology-based Conceptual Vectors." In: *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science (NLPCS'08).* Barcelona, Spain 2008, pp. 153–158.

[15] R. Mihalcea. "Knowledge-Based Methods for WSD." In: *Word Sense Disambiguation: Algorithms and Applications.* Ed. by E. Agirre and P. Edmonds. Dordrecht, the Netherlands: Springer, 2006. Chap. 5, pp. 107–132.

[16] H. T. Ng, B. Wang, and Y. S. Chan. "Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study." In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics.* Sapporo, Japan 2003, pp. 455–462.

[17] P. Vossen. "EuroWordNet: a multilingual database for information retrieval." In: *In Proceedings of the DELOS Workshop on Cross-language Information Retrieval.* 1997, pp. 5–7.

[18] R. Zajac. "Structuring a Multilingual Multipurpose Lexical Database Using a Simple Interlingual Approach." In: *Proceedings of AMTA-96 Workshop on Interlinguas.* Montreal, Canada 1996.