Cognate Mapping — A Heuristic Strategy for the Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon

Stefan Schulz[a,b] Kornél Markó[b,c] Eduardo Sbrissia[a] Percy Nohama[a] Udo Hahn[c]

[a]Master Program in Health Technology, Paraná Catholic University, Curitiba, Brazil

[b]Department of Medical Informatics, Freiburg University Hospital, Germany

[c]Computational Linguistics Research Group, Jena University, Germany

http://www.coling.uni-freiburg.de/

Abstract

We deal with the automated acquisition of a Spanish medical subword lexicon from an already existing Portuguese seed lexicon. Using two non-parallel monolingual corpora we determined Spanish lexeme candidates from Portuguese seed lexicon entries by heuristic cognate mapping. We validated the emergent lexical translation hypotheses by determining the similarity of fixed-window context vectors on the basis of Portuguese and Spanish text corpora.

1 Introduction

Medical language presents a unique combination of challenges for language engineering, with a focus on applications such as information retrieval, text mining and information extraction. Document collections – on the Web or in clinical databases – are usually very *large* and *dynamic*. In addition, medical document collections are truly *multilingual*. Furthermore, the user population which access medical documents are really *diverse*, ranging from physicians and nurses to laypersons, who use different jargons and sublanguages. Therefore, the simplicity of the content representation of the documents, as well as automatically performed intra- and interlingual lexical mappings or transformations of equivalent expressions, become crucial issues for an adequate machine support.

We respond to these challenges in terms of the MORPHOSAURUS system (an acronym for MORPHeme TheSAURUS). It is centered around a new type of lexicon, in which the entries are subwords, i.e., semantically minimal, morpheme-style units (Schulz and Hahn, 2000). Intralingual as well as interlingual synonymy is then expressed by the assignment of subwords to concept-like equivalence classes. As subword equivalence classes abstract away from subtle particularities within and between languages, and reference to them is achieved via a language-independent code system, they form an interlingua characterized by semantic identifiers. Compared to relationally richer, e.g., WORDNET based, interlinguas as applied for cross-language information retrieval (CLIR) (Gonzalo et al., 1999; Ruiz et al., 1999), we use a rather limited set of semantic relations and pursue a more restrictive approach to synonymy. In particular, we restrict ourselves to the specific sublanguage used in the context of the medical domain. Our claim that this interlingual approach is useful for the purpose of cross-lingual text retrieval and categorization has already been experimentally supported (Schulz et al., 2002; Markó et al., 2003).

The quality of cross-lingual indexing fundamentally depends on the underlying lexicon and thesaurus. Its manual construction and maintenance is costly and error-prone. Therefore, machine-supported lexical acquisition techniques increasingly deserve attention. Whereas in the medical domain parallel corpora are only available for a limited number of language pairs, unrelated (i.e., non-parallel, non-aligned) corpora might provide sufficient evidence for cognate identification, at least in languages which are closely related.

In this paper, we present the results of such an experiment. We have chosen Spanish and Portuguese as a pair of closely related languages. Both languages exhibit a high degree of similarity in their lexical inventory, as well as in the rules governing word formation. Accordingly, a Portuguese native speaker is able to understand technical texts in Spanish without much effort, and vice versa. In both languages, there is also an increasing number of electronic texts available, so that a cross-lingual search interface would significantly improve the accessibility of domain relevant documents.

2 Lexicographic Aspects of Morpho-Semantic Indexing

We briefly outline the lexicographic and semantic aspects of our approach, called *Morpho-Semantic Indexing* (henceforth, MSI), which translates source documents (and queries) into an interlingual representation in which their content is represented by language-independent semantic descriptors.

2.1 Subwords as Lexicon Units

Our work is based on the assumption that neither fully inflected nor automatically stemmed words constitute the appropriate granularity level for lexicalized content description. Especially in scientific and technical sublanguages, we observe a high frequency of domain-specific and content-bearing suffixes (e.g., *'-itis'*, *'-ectomia'* in the medical domain), as well as the tendency to construct utterly complex word forms such as *'pseudo⊕hypo⊕para⊕thyroid⊕ism'*, *'gluco⊕corticoid⊕s'*, or *'pancreat⊕itis'*.[1] In order to properly account for the particularities of "medical" morphology, we introduced subwords (Schulz et al., 2002) as self-contained, semantically minimal units and motivated their existence by their usefulness for document retrieval rather than by linguistic arguments.

The minimality criterion is quite difficult to define in a general way, but its implications can be illustrated by the following example. Given the text token *'diaphysis'*, a linguistically plausible morpheme decomposition would possibly lead to *'dia⊕phys⊕is'*. From a medical perspective, a segmentation into *'diaphys⊕is'* seems much more reasonable, because the linguistically canonical morphological decomposition is far too fine-grained and likely to create too many ambiguities. For instance, comparable 'low-level' segmentations of semantically unrelated tokens such as *'***dia***⊕lyt⊕ic'*, *'***phys***⊕io⊕logy'* lead to morpheme-style units *'dia'* and *'phys'*, which unwarrantedly match segmentations such as *'***dia***⊕***phys***⊕is'*, too. The (semantic) self-containedness of the chosen subword is often supported by the existence of a synonym, e.g., for *'diaphys'* we have *'shaft'*.

2.2 Subword Lexicon and Thesaurus

Subwords are assembled in a multilingual lexicon and thesaurus, which contain subword entries, special subword attributes and semantic relations between subwords. Up until now, the lexicon and the thesaurus have both been constructed manually, with the following considerations in mind:

- Subwords are entered, together with their attributes such as language (English, German, Portuguese) and subword type (stem, prefix, suffix, invariant). Each lexicon entry is assigned a unique identifier representing one synonymy class, the MORPHOSAURUS identifier (MID), which contains this entry as its unique member.
- Synonymy classes which contain intralingual synonyms and interlingual translations of subwords are fused. We restrict intra- and interlingual semantic equivalence to the context of medicine.
- Semantic links between synonymy classes are added. We subscribe to a shallow approach in which semantic relations are restricted to a paradigmatic relation *has-meaning*, which relates one ambiguous class to its specific readings,[2] and a syntagmatic relation *expands-to*, which consists of predefined segmentations in case of utterly short subwords.[3]

We refrain from introducing hierarchical relations between MIDs, because such links can be acquired from domain-specific vocabularies, e.g., the Medical Subject Headings (MESH, 2001).

Table 1 depicts how source documents (cf. the first column with an English and Portuguese fragment) are converted into an interlingual representation by a three-step procedure. First, each input word is orthographically normalized in terms of lower case characters and according to language-specific rules for the transcription of diacritics (second column). Next, words are segmented into sequences of semantically plausible sublexical items according to the subwords listed in the lexicon (third column). Finally, each meaning-bearing subword is replaced by its language-independent semantic identifier, the MID, which unifies intralingual and interlingual (quasi-)synonyms. Then, the system yields the interlingual output representation of the system (fourth column).

The manual construction of the trilingual subword lexicon and the subword thesaurus has consumed, up until now, three and a half person years. The project originally started from a bilingual German-English lexicon, while the Portuguese part was added in a later project phase. The combined subword lexicon contains 58,479 entries,[4] with 21,397 for English, 22,053 for German, and 15,029 for Portuguese.

Taking into account, on the one hand, the outstanding importance of Spanish as a major Western

[1] '⊕' denotes the concatenation operator.

[2] For instance, {*head*} ⇒ {*zephal,kopf,caput,cephal,cabec, cefal*} *OR* {*leader,boss,lider,chefe*}

[3] For instance, {*myalg*} ⇒ {*muscle,muskel,muscul*} ⊕ {*schmerz, pain,dor*}

[4] Just for comparison, the size of WORDNET assembling the lexemes of general English in the 2.0 version is on the order of 152,000 entries (`http://www.cogsci.princeton.edu/~wn/doc.shtml`, last visited on January 3, 2004). Linguistically speaking, the entries are basic forms of verbs, nouns, adjectives and adverbs.

Original Document	**Orthographic Normalization**	**Morphological Segmentation**	**Semantic Normalization**
High TSH values suggest the diagnosis of primary hypothyroidism while a suppressed TSH level suggests hyperthyroidism.	high tsh values suggest the diagnosis of primary hypothyroidism while a suppressed tsh level suggests hyperthyroidism.	high tsh value s suggest the diagnos is of primar y hypo thyroid ism while a suppress ed tsh level suggest s hyper thyroid ism.	**#up# tsh #value# #suggest# #diagnost# #primar# #small# #thyre# #suppress# tsh #nivell# #suggest# #up# #thyre#** .
A presença de valores elevados de TSH sugere o diagn´ostico de hipotireoidismo prim´ario, enquanto n´ıveis suprimidos de TSH sugerem hipertireoidismo.	a presenca de valores elevados de tsh sugere o diagnostico de hipotireoidismo primario, enquanto niveis suprimidos de tsh sugerem hipertireoidismo.	a presenc a de valor es elevad os de tsh suger e o diagnost ico de hipo tireoid ismo primari o, enquanto niveis suprimid os de tsh suger em hiper tireoid ismo.	#actual# **#value# #up# tsh #suggest# #diagnost# #small# #thyre# #primar#** , **#nivell# #suppress# tsh #suggest# #up# #thyre#** .

Table 1: Morpho-Semantic Indexing Example for English (row 1) and Portuguese (row 2): The original document (column 1) is orthographically transformed (column 2), segmented according to the subword lexicon (column 3), while content-bearing subwords are mapped to MSI-specific equivalence classes whose identifiers (MIDs) are automatically generated by the system (column 4). (Bold MIDs co-occur in both documents.)

language and, on the other hand, the close lexical ties between Portuguese and Spanish as Romance languages, we intended to augment the existing MORPHOSAURUS system by Spanish as its fourth language and at the same time reuse the knowledge of Portuguese for the purpose of speeding up and facilitating the Spanish lexicon acquisition.

3 Experiments

We use the following resources for the experiments:

- A Portuguese subword lexicon, as described in the previous section.
- A manually created list of 842 Spanish affixes.
- Medical corpora for Spanish and Portuguese. These corpora were compiled exploiting heterogeneous WWW sources. The size of the acquired corpora amounts to 2,267,841 tokens with 118,021 types for Spanish and 3,406,589 tokens with 133,146 types for Portuguese.
- Word frequency lists generated from these corpora, for Spanish and Portuguese.

3.1 Spanish Subword Generation

In order to acquire a first-shot Spanish subword lexicon we designed the following lexeme generation strategy: Using the Portuguese lexicon, identical and similarly spelled Spanish subword candidates (cognates) are generated. As an example, the Portuguese word stem *'estomag' ('stomach')* is identical with its Spanish cognate. An example for a pair of similar stems is *'mulher' ('woman')* (Portuguese) vs. *'mujer'* (Spanish). Similar subword candidates

Rule (P → S)	Portuguese Example	Spanish Example
qua → cua	quadr	cuadr
eia → ena	veia	vena
ss → s	fracass	fracas
lh → j	mulher	mujer
lh → ll	detalh	detall
l → ll	lev	llev
i → y	ensai	ensay
f → h	formig	hormig
+ca → za	cabeca	cabeza
+o+ → ue	sort	suert
...	...	...

Table 2: Sample of Portuguese-to-Spanish String Substitution Rules

were generated by applying a set of string substitution rules some of which are listed in Table 2. In total, we formulated 45 rules as a result of identifying common-language Portuguese-Spanish cognates in a commercial dictionary. Some of these substitution patterns cannot be applied to starting or ending sequences of characters in the Portuguese source subword. These regularities are captured by using a wildcard ('+' in Table 2) representing at least one arbitrary character.

First, for each Portuguese lexicon entry (n = 14,183 stems and invariants, excluding affixes), all possible Spanish variant strings were generated based upon the set of string substitution rules. This led, on the average, to 9.53 Spanish variant hypotheses per Portuguese subword entry (ranging from 5.3 variants for high-frequency four-character words to 355.2 for low-frequency 17-character words). All

these candidates were subsequently compared to the Spanish word frequency list, we had previously compiled from our Spanish text corpus. Wherever a left-sided string match (in the case of stems) or an exact one (in the case of invariants) occurred, the matching string was listed as a potential Spanish cognate of the Portuguese subword it originated from. Whenever several Spanish substitution alternatives for a Portuguese subword had to be considered (cognate ambiguity) that particular one was chosen which had the closest relative distribution in the corpus-derived Spanish word frequency list, when compared to its Portuguese equivalent in the Portuguese word list. As a result, we obtained a list of tentative Spanish subwords each linked by the associated MIDs to its corresponding cognate in the Portuguese lexicon.

Quantitatively, starting from 14,183 Portuguese subwords, a total of 132,576 Spanish subword candidates were created using the string substitution rules. Matching these Spanish candidates against the Spanish corpus and allowing for a maximum of one Spanish candidate per Portuguese subword, we identified 11,206 tentative Spanish cognates (79% of the Portuguese seed lexicon) which are linked to a total of 8,992 MIDs from their Portuguese correlates (hence, 2214 synonym relationships have also been hypothesized). 2,977 generated items could not be found in the Spanish corpus, at all.

3.2 Manual Semantic Validation

One of the authors evaluated manually a random sample of 388 (3.5% of all generated) cognate pairs in order to identify *false friends*, i.e., similar words in different languages with different meanings. In our sample we found, e.g., the Spanish candidate *'*crianz*' for the Portuguese '*crianc*' (the normalized stem of '*criança*'; English: '*child*'). The correct translation of Portuguese '*crianc*' to Spanish, however, would have been '*nin*' (the stem of '*niño*'), whilst the Spanish '*crianz*' refers to '*criac*' (stem of '*criação*' in Portuguese; English: '*breed*'). Taking these false friend errors into account, the automatic generation of Portuguese-Spanish cognate pairs still yields 89,4% accuracy.

Assuming then that approximately 1,188 false friends are among the list of 11,206 generated Spanish subword translations (10.6%), the question arises how to distinguish false friends from true positives (cognates). Because a manual examination of the entire candidate set is a tedious and still error-prone work, we shifted our attention to automatic semantic validation techniques.

3.3 Automatic Semantic Validation

In order to automatically validate all the generated cognate pairs, we examined the local context in which these cognates occur in *non-parallel* corpora of both languages involved. The basic idea that underlies this approach is that a subword that appears in a certain context should have a (true positive) cognate that occurs in a similar context, at least when (*very*) large corpora are taken into account. Cognate similarity can then be measured in terms of context vector comparison (cf. also Rapp (1999) or Koehn and Knight (2002)).

We therefore processed the Portuguese corpus using the morpho-semantic normalization routines as discussed in Section 2. In the next step, we created a context vector for each MID, the components of which contained the relative frequencies of co-occurring MIDs in a local window of four subsequent, yet unordered MID units (a size also endorsed by Rapp (1999)).

In order to compute the context vector for each Spanish subword candidate, we then constructed a seed lexicon with all the automatically created Spanish subword candidates, together with the list of Spanish affixes. Based on this lexicon, the Spanish corpus was morphologically normalized in the same way, using the MIDs that were licensed by the Portuguese cognates. For each of the candidate cognate MIDs, we built a corresponding context vector.

We then measured the context similarity for each MID considering its Portuguese source context and the corresponding Spanish one. We chose two similarity metrics, *viz.* the well-known cosine metric (Salton and McGill, 1983) and an inverted, normalized (within the interval [0,1]) variant of the city-block metric (advocated by Rapp (1999) as an alternative that outperformed cosine in his experiments).

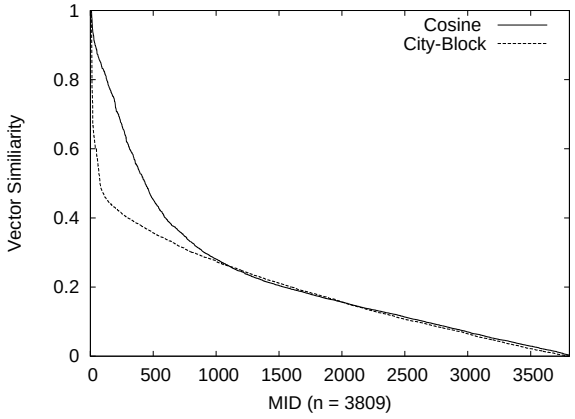

Figure 1: Context Similarity of MIDs Representing Portuguese and Spanish Sources

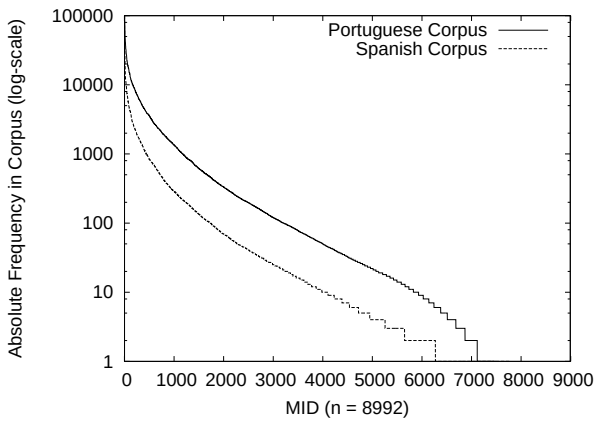

Figure 2: Distribution of MIDs in the Portuguese and Spanish Corpora

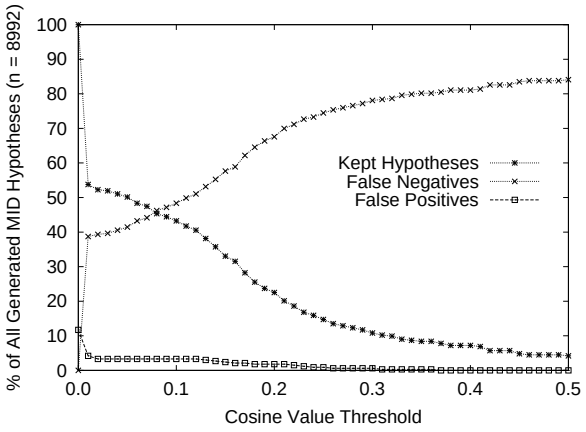

Figure 3: The Effects of Applying a Threshold Value to the Cosine Metrics for the Validation of MID Hypotheses

Figure 1 depicts the resulting curves. Both metrics reveal almost the same characteristics. Only for higher similarities, city-block allows a more fine-grained distinction.

For 5,183 (57.6%) from 8,992 pairs of MIDs (one from a 'Portuguese' vector, the other from a 'Spanish' vector), no vector similarity at all could be measured. We distinguish between the following cases:

- There was no MID occurrence in the Spanish corpus.
- There was a MID occurrence in the Spanish corpus, but none in the Portuguese one.
- The vectors were orthogonal, i.e., the contexts did not overlap at all, although the MID occurred in the Spanish corpus, as well as in the Portuguese one. This can be interpreted in two ways: For reasonably frequent MIDs (cf. Figure 2 for the distribution in the corpora) this is the strongest evidence for false friends (formal cognates which are not semantically related), whereas for sparsely distributed MIDs, it does hardly permit any valid judgment concerning their status as false or true cognates.

On the other hand, 1,540 MID pairs (in the sense from above) exceed similarity values of 0.2 (17.1%) and 2,065 pairs still share values greater than 0.15 (23%). The obvious question is: What is an adequate threshold?

Figures 3 and 4 convey an answer to this question. Both figures are meant to illustrate the trade-off when one increases the threshold for the similarity of both vectors, the Portuguese and the Spanish one, for the MIDs under consideration. The central notion in these two figures is that of *Kept Hypotheses*, i.e., the proportion of MIDs for which

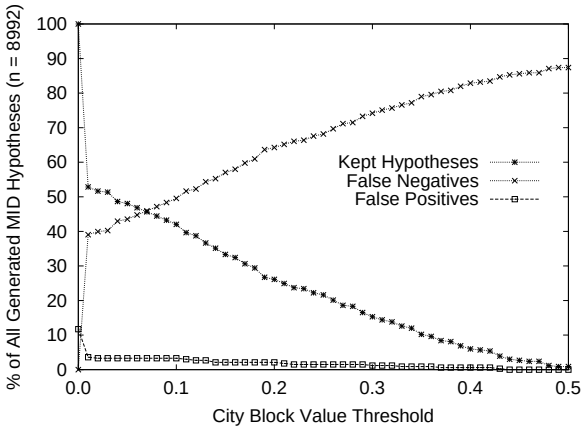

Figure 4: The Effects of Applying a Threshold Value to the City-Block Metrics for the Validation of MID Hypotheses

the assignment of the underlying cognate is judged as being semantically valid. When we consider all (100%) of the generated MIDs (n=8,992) as valid (hence, cosine and city-block are both zero), we get 953 false positives (given our empirically determined accuracy rate of 89.4%, and, hence, error rate of 10.6%) and, obviously, no false negative. Alternatively, when we consider instead 50% of the generated MIDs (n=4,496) as valid (with thresholds for cosine set at 0.05 and for city-block at 0.035), we get 297 (3.3%) false positives, and the number of false negatives increases at a level of 3,687 (around 41%, for both metrics). In order to reduce the set of false friends to zero using the cosine metric, 92.2% of all generated MID cognates will be rejected by the automatic validation for manual revision (analogously, the number of false negatives will increase). Interestingly, the same procedure using the city-block metric will lead to a rejection rate of 97%.

At a first glance, this seems to contradict the statement of Rapp (1999), who found in a number of experiments that the city-block metric yields the best results among others, *viz.* cosine and Jaccard measure, Euclidean distance and scalar product. However, his measures were taken to *find the most similar vector for a given word* in order to automatically identify word translations. On the other hand, in our experiments, we intended to express the *degree of similarity given a pair of cognates*. We hypothesized that the city-block metric allows a more fine grained similarity judgment whilst others, e.g., cosine, the Jaccard and Dice coefficient, etc., which only account for overlapping elements of a vector, have a stronger demarcation power.

Summarizing, when we increase the similarity thresholds, the number of MID hypotheses decreases as does the number of false positives (already at a rather low level), while the number of false negatives increases almost inversely related to the number of MID hypotheses. Therefore, it is up to the lexicon engineer to determine the level of preselection in these three dimensions. We also conclude from our experiments that a much larger corpus is needed in order to collect reasonable context evidence for the infrequent MIDs, in particular.

4 Related Work

The rise of the empirical paradigm in the field of machine translation is, to a large degree, due to the wide-spread availability of parallel corpora (Brown et al., 1990). They also constitute an important resource for the automated acquisition of translational lexicons (Turcato, 1998). Unfortunately, the limited availability of parallel corpora (e.g., the Canadian *Hansard* corpus of English and French parliament debates) restricts this method to a few language pairs, mostly focused on specific sublanguages (e.g. politics, legislation, economy). Neither exist such parallel corpora for the medical sublanguage, nor for the particular language pair, Spanish and Portuguese, we focus on in this work.

The acquisition of unrelated, albeit comparable corpora is much easier. Rapp (1999) used unrelated parallel corpora in order to learn English and German word-to-word translations. His approach is based on similarity measures and context clues, using a seed lexicon of trusted translations. Koehn and Knight (2002) derived such a seed lexicon from German-English cognates which were selected by using string similarity criteria. An additional boost can be achieved by retrieving content-related document pairs using CLIR techniques (Utsuro et al., 2003). An alternative generative approach is proposed by Barker and Sutcliffe (2000) who created Polish cognate candidates out of an English wordlist using a set of string mapping rules.

Pirkola et al. (2003) used aligned translation dictionaries as source data. Based on that, they created an algorithm to automatically generate transformation rules from five different languages to English, including Spanish. Applying a two-step technique (translation rules and fuzzy n-gram matching), they achieved 81.1% of average precision in a Spanish-to-English context covering biomedical words only. However, their evaluation metrics considerably differed from ours, since they considered multiple hypotheses.

Our work differs from these precursors in many ways. First of all, due to domain and language restrictions the size of our corpora is much smaller than the commonly used newspaper corpora. For the same reasons, CLIR techniques for retrieving comparable documents are not yet available (on the contrary, the goal of our work is to provide resources for a medical CLIR system). Thirdly, the two languages are so similar that a high amount of translations could already be acquired by applying string mapping rules (this approach to cognate mapping has also been discussed by MacWhinney (1995) for second language acquisition of human learners). Finally, rather than acquiring bilateral word translation, our focus lies on assigning subwords to interlingual semantic identifiers.

5 Conclusions and Further Work

In a first round of experiments, we have shown that a considerable amount of Portuguese subwords from the medical domain could be mapped to Spanish cognate stems applying simple string transformation rules. We then used the local context in language-specific corpora in order to validate these cognate pairs. However, our results also reveal the limitations of such an approach, at least for infrequent stems, due to the small corpus size. Accordingly, for future experiments one has to provide much larger text corpora, paticularly in the next steps of our experiments, in which the Spanish lexicon will be completed by subwords which can*not* be generated from their Portuguese translations. Here, we will acquire new Spanish lexeme candidates by automated stemming, and retrieve their Portuguese translations by exploring their local context. This requires, however, huge corpora, exceeding the current ones by several orders of magnitude. Additionally, their documents will have to be related using clustering techniques. The usability of the resulting, mainly automatically generated Spanish extension

of the MORPHOSAURUS lexicon for the purpose of cross-language text retrieval can then be evaluated in real CLIR experiments as previously done for English, German and Portuguese (cf. Hahn et al. (2004)).

Acknowledgements.
This work was partly supported by the German Research Foundation (DFG), grant KL 640/5-1, and by the Brazilian National Council for Scientific Research and Development (CNPq), grants 551277/01-7 and 550240/03-9.

References

Gosia Barker and Richard F. E. Sutcliffe. 2000. An experiment in the semi-automatic identification of false-cognates between English and Polish. In *Proceedings of the Irish Conference on Artificial Intelligence and Cognitive Science*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Julio Gonzalo, Felisa Verdejo, and Irina Chugur. 1999. Using EUROWORDNET in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence*, 13(7):647–678.

Udo Hahn, Kornél Markó, Michael Poprat, Stefan Schulz, and Joachim Wermter. 2004. Crossing languages in text retrieval via an interlingua. In *Proceedings of the 7th International RIAO'04 Conference*, pages 82–99.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Unsupervised Lexical Acquisition: Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9–16. Association for Computational Linguistics.

Brian MacWhinney. 1995. Language-specific prediction in foreign language learning. *Language Testing*, 12(3):292–320.

Kornél Markó, Phillip Daumke, Stefan Schulz, and Udo Hahn. 2003. Cross-language MESH indexing using morpho-semantic normalization. In *AMIA'03 – Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association*, pages 425–429. Philadelphia, PA: Hanley & Belfus.

Ari Pirkola, Jarmo Toivonen, Heikki Keskustalo, Kari Visala, and Kalervo Järvelin. 2003. Fuzzy translation of cross-lingual spelling variants. In *SIGIR 2003 – Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 345–352. Toronto, Canada, 2003, New York, NY: ACM.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 519–526. San Francisco, CA: Morgan Kaufmann.

Miguel Ruiz, Anne Diekema, and Páraic Sheridan. 1999. CINDOR conceptual interlingua document retrieval: TREC-8 evaluation. In *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 597–606. National Institute of Standards and Technology (NIST). NIST Special Publication, No. 500-246.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York, NY: McGraw Hill.

MESH. 2001. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

Stefan Schulz and Udo Hahn. 2000. Morpheme-based, cross-lingual indexing for medical document retrieval. *International Journal of Medical Informatics*, 59(3):87–99.

Stefan Schulz, Martin Honeck, and Udo Hahn. 2002. Biomedical text retrieval in languages with a complex morphology. In *Proceedings of the ACL 2002 Workshop 'Natural Language Processing in the Biomedical Domain'*, pages 61–68. New Brunswick, NJ: Association for Computational Linguistics (ACL).

Davide Turcato. 1998. Automaticaly creating bilingual lexicons for machine translation from bilingual text. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, volume 2, pages 1299–1306. San Francisco, CA: Morgan Kaufmann.

Takehito Utsuro, Takashi Horiuchi, Takeshi Hamamoto, Kohei Hino, and Takeaki Nakayama. 2003. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In *EACL'03 – Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 355–362. Association for Computational Linguistics.