Simple Preposition Correspondence: A problem in English to Indian language Machine Translation

Samar Husain, Dipti Misra Sharma, Manohar Reddy
{samar@research.iiit.net,dipti@mail.iiit.net,
manohar@students.iiit.net}
Language Technologies Research Centre,
IIIT, Hyderabad, India.

Abstract

The paper describes an approach to automatically select from Indian Language the appropriate lexical correspondence of English simple preposition. The paper describes this task from a Machine Translation (MT) perspective. We use the properties of the head and complement of the preposition to select the appropriate sense in the target language. We later show that the results obtained from this approach are promising.

1 Introduction

The task of identifying the appropriate sense from some target language (here, Hindi and Telugu) for a given simple preposition in some source language (here, English) is rather complex for an MT system, and noting that most foreign language learners are never able to get a firm hold on prepositions of a new language (Brala, 2000), this should not be surprising. A simple example illustrates the problem:

(1a) *He bought a shirt* ***with*** *tiny collars.*
'with' gets translated to *vaalii* in Hindi (hnd).
and as *kaligi unna* in Telugu (tlg).
(1b) *He washed a shirt* ***with*** *soap.*
'with' gets translated to *se* in hnd.
and as *to* (suffixed to head noun) in tlg.

For the above English sentences, if we try to swap the senses of '*with*' in their corresponding target translation, the resulting sentences either become ill-formed or unfaithful to their English source. The pervasive use of preposition (or its equivalent in a given language) in most of the languages makes it a crucial element during translation. Inappropriate sense selection of a preposition during machine translation can have a negative impact on the quality of the translation, sometimes changing the semantics of the sentence drastically, thereby making the preposition sense selection module a critical component of any reliable MT system.

Finding the proper attachment site for the preposition in English, i.e. getting the correct parse for the prepositional phrase (PP) is a classic problem in MT, and this information can be used to identify the sense of a preposition. Figure 1 and Figure 2 below show the correct attachment site of PPs in example (1a) and (1b) respectively.

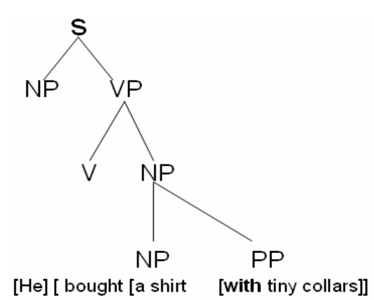

Figure 1

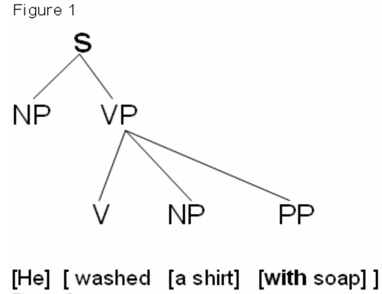

Figure 2

Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions, pages 51–58,
Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

The correct parse of the PP helps us in selecting the appropriate sense. However, finding the appropriate attachment only reduces the problem. It does not lead to a 'complete solution'. The following examples (2a, 2b and 3a, 3b) have the same attachment site but take different senses in the target language:

(2a) *He has had fever* ***for*** *two days now.*
'for' gets translated as *se* in hnd.
and as *nundi* in tlg.
(2b) *He had fever* ***for*** *two days.*
'for' gets translated as *taka* in hnd.
Not translated in tlg.

(3a) *He is going* ***to*** *Delhi.*
'to' gets translated as *ko*, or preferably left un-translated in hnd.
and in tlg as *ki* (suffixed to the head noun), or may be left un-translated.
(3b) *He is going* ***to*** *his mother.*
'to' gets translated as *ke paasa* in hnd.
and *daggaraku* in tlg

After looking at cases such as (2a), (2b) and (3a), (3b) where the parse is same i.e., preposition 'for and 'to' get attached to the main verb 'have' and 'go' respectively, it is clear that we need to come up with some criterion which can help us in achieving our task.

There has been extensive work on understanding prepositions linguistically, often from various angles. Syntactically (Jackendoff, 1977; Emonds, 1985; Rauh, 1993; Pullum and Huddleton, 2002), from a Cognitive perspective (Lakoff and Johnson, 1980; Langacker, 1987; Brala, 2000), Semantically by (Saint-Dizier and Vazquez, 2001; Saint-Dizier, 2005), and the Pragmatic aspects by (Fauconnier, 1994).

The work of automatically selecting the correct sense has also received good amount of attention and there have been many attempts to solve the problem. (Japkowicz et. al, 1991) attempts to translate locative prepositions between English and French. The paper introduces the notion of 'representation of conceptualization' based in turn on (Grimaud, 1988). The paper synthesizes this idea with the thesis of ideal meaning (Herskovits, 1986). (Tezuka et. al, 2001) have tried to resolve conceptual geographical prepositions using inference rule based on cognitive maps which people have of the external world. (Hartrumpf et al., 2005) use knowledge representation formalism for PP interpretation.

Some studies pertain to systems which have been implemented for MT; (Gustavii, 2005) uses aligned parallel corpora to induce automatic rules by applying transformation-based learning. (Alam, 2004) make use of contextual information to determine the meanings of *over*. (Trujillo, 1992) use a transfer rule based approach to translate locative PP-phrase, the approach uses the dependency relations marked as indices with individual word and a bilingual lexicon which has mapping between source and target lexical item (with indices). (Naskar and Bandyopadhyay, 2005) look at the semantics of the head noun of the reference object (this is their main criterion) to get the lexical meaning of prepositions in an English-Bengali MT system.

The current paper presents a study of prepositions at, for, in, on, to and with in context of English to Indian language MT system. The paper is arranged as follows; Section 2 describes our approach to solving the mentioned task, the 3rd section shows the performance of our approach along with the error analysis during the testing phase, we conclude the paper along with some future direction in section 4.

2 Our Approach

All the previous attempts can be broadly classified into 3 main categories; *one*, where the preposition is the main focus, concentration is on the semantics (cognitive or lexical) of the preposition; *second*, focus on the verb and the PP which the verb takes as argument; and *lastly*, the head noun of the PP becomes the deciding factor to get the appropriate sense.

Very few approaches, like (Alam, 2004; Saint-Dizier and Vazquez, 2001), consider both, the head (modified) and the complement (modifier) information, to decide the sense of the preposition. The modified (or head) is the head of the phrase to which the PP attaches. The modifier (or complement) is the head noun of the PP. The following examples show very clearly why given a preposition we cannot depend only on the modified or the modifier separately, and that we must consider them both to solve the problem.

Considering only the modifier (the complement);

(4a) *He apologized* ***to*** *his mother.*
'to' gets translated as *se* in hnd
& *ki* (suffixed to the head noun) in tlg
(4b) *He went* ***to*** *his mother.*
'to' gets translated as *ke paasa* in hnd
& as *daggaraku* in tlg

Considering only the modified (the head);

(5a) *He waits for her* ***at*** *night.*
'at' gets translated as *meM* in hnd
& not translated in tlg
(5b) *He waits for her* ***at*** *the station.*
'at' gets translated as *par*
& as *lo* in tlg

Only considering the modifer *'his mother'* in 4a and 4b is not sufficient, likewise taking only the modified *'waits'* in 5a and 5b will be insufficient, both the pairs take different senses and have the same partial contextual enviornment which is misleading. Hence, the combined context of complement-head forms a better candidate for solving the problem. We come across plenty of cases where isolated information of modifier/modified can be misleading.

The task of preposition sense selection can be divided into;

(a) Getting the correct parse (the task of PP attachment, identification of phrasal verb, etc.),
(b) Context and semantic extraction,
(c) Sense selection.

This paper describes the algorithm for achieving the above mentioned steps. *We assume the input to our module has the correct parse, i.e. Step (a) above is assumed here.* The proposed algorithm is a component in English to Indian language MT system[1], therefore, the required input can be presumed to be available. Steps (b, c) above are rule based, which make use of the modifier-modified relation, these relations and the properties of modifier/modified form the core of the context in step (b). We then apply a series of rules, which specify the context and semantics in which a sense is expected to occur.

2.1 Context and semantic extraction

Extraction of context and semantic information (of modifier/modified) is done automatically by various sub-modules which are combined together to perform the overall task. We use the word '*context'* very loosely. A context for us is a combination of various properties which can be syntactic or lexical, or both; syntactic context can be modifier-modified relation, lexical properties can be morphological information such as TAM (tense, aspect and modality) of a verb, class of the verb (Levin, 1993), category of the lexical item and in some cases the lexical item itself.

The semantics of the modifier and the modified are captured using WordNet (Miller, 1990), and certain other resources such as person, place dictionaries, place and time filters (these filters make use of syntactic cues to mark basic time and place), etc. We use WordNet to get the hypernyms of a word. By using this property we can easily get the broader, more general class/concept for a modifier/modified. Although effective and very intuitive, this method has its own problems. We will elaborate these problems in section 3.2. WordNet is also used to identify person and place names by using the hyponym tree for person and place.

Along with the WordNet, as mentioned above, we use certain other filters such as place and time. They are used prior to using WordNet. In case a rule requires the modifier to be a *place* (rules are explained in 2.2), this information is acquired from the place filter. If the filter's result is negative we use WordNet. Dictionaries and POS tags are checked for identifying proper names, we use a proper name dictionary as POS taggers tend to have a fixed upper limit especially when it comes to the identification of named entities. In essence, the linguistic resources are used in the following order;

(1) Dictionaries,
(2) Time & Place filter,
(3) WordNet.

Preliminary results have shown that certain prepositions occurring in the PP complement of certain verb classes (Levin, 1993) translate to a specific sense in Hindi. For example, preposition 'at' in the case of *peer* verbs always translates to *kii tarapha* or *kii ora* in Hindi. This knowledge can

[1] (http://shakti.iiit.ac.in). Note here that the proposed algorithm has been tested with Shakti version 0.83x which has still not been released. The released version is 0.73.

be very informational and we plan to pursue this aspect in the future.

2.2 Sense Selection

We have noticed in the previous examples that the prepositions from English either get translated as suffixes to the head noun of the PP (in Telugu) or as postpositions (in Hindi and Telugu). An example where a preposition in English gets translated as postposition in its Telugu translation is shown below;

(6) *The book is* ***on*** *the table.*
‘buka taibila ***paiina*** undi’
‘Book’ ‘table’ ‘on’ ‘there’

We select the correct sense of the preposition based on a series of rules which are applied linearly. These rules have been manually constructed. We have tried to make the rules mutually exclusive, so that there are no clashes. Also, by making sure that the rules are mutually exclusive we don’t need to worry about the order in which the rules are listed out in the rule file, thus making the rule file less fragile. These rules currently cover around 20 high frequency English prepositions, these prepositions vary in their degree of ambiguity; some are highly ambiguous (e.g. to, by, with, etc.), whereas some are less ambiguous (e.g. against, around, as, etc.), hence these are easier to handle.

Various senses on the target side for a given English preposition are selected on the basis of rules listed out in a file. The rule file comprises of tuples, each having 6 attributes.

The attributes are listed below;

a) Source Language preposition
b) Modified category
c) Constraints on the modified item
d) Modifier category
e) Constraints on the modifier item
f) Dictionary sense id of the source language preposition

An example of a tuple:
 # at, v, -, n, place_close, at%p%5

(7) *He has opened a school* ***at*** *his home.*
‘usane apne ghara ***mem*** eka skuula kholaa hei’
‘He *erg*’ ‘his’ ‘house’ ‘at’ ‘one’ ‘school’ ‘open’ ‘is’

The rule above requires the modifier to be a noun and places a constraint “place_close” on it. We map this constraint (place_close) with some set of lexical items found in a synset of a hypernym obtained from WordNet. For example, “place_close” might correspond to ‘housing’, ‘lodging’, ‘building’, etc in a synset. In essence “place_close” is place holder for different relations which might be present in a synset. The modified category and the modifier category can be extracted after the correct parse of the PP is known; the constraints applied on the modified and modifier item (point c, e above) can be of various kinds, some of them are;

- Semantic relations corresponding to WordNet hypernyms for a given word
- Presence of the lexical item in some list (eg. verb class)
- Semantic property such as ‘time’ or ‘place’
- Lexical property such as aspect, negativity etc.

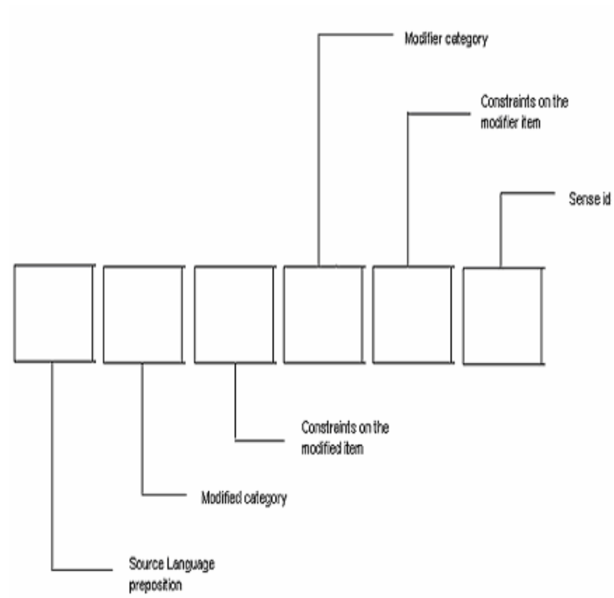

Figure 3: Single rule tuple

The constraints specified in a tuple can be combined together using logical operators such as ‘and’, ‘or’, ‘negation’. So, for a single rule, multiple constraints can be introduced. For a sense, if needed, complex constraints can be introduced which must be satisfied.

#for, v, L[2]:for.dat && *aspect:continuous, n, time, for%p%5*

(8) *He has been playing* ***for*** *years.*
'vaha kaii saalo ***se*** khela rahaa hai '
'He' 'many' 'years' 'for' 'play' 'cont.' 'is'

The above rule (for the Hindi translation) has two constraints for the modified (which is a verb in this case), the two constraints have been combined using an 'and' operator (represented using two ampersands, '&&'). Only if the two constraints are satisfied, the constraint is considered as satisfied else it is considered as failed. The use of different logical operator gives a lot of expressive power to a single rule. Sometimes it might be desirable to place multiple constraints together, because for a given sense these constraints always occur together, and by listing them as separate rules we will miss out the fact that they co-occur.

It is not always necessary (or possible) to fill the constraint fields. In fact, sometimes it is even desirable to leave them unspecified. In such a case we place a hyphen in that field, such as the following rule;

 # at, v, -, n, place_close, at%p%5

In the above rule, the constraint for the modified field is unspecified. There are also cases when it is not desirable to have a translated preposition corresponding to its source;

 # to, L: verbs.txt, -, n, place, ZZ

(9) *He went* ***to*** *Delhi.*
'vaha dilli gayaa' (in hnd)
'He' 'Delhi' 'went'

The 'ZZ' in the above rule signifies that the translated sentence will have no preposition corresponding to the preposition 'to' when it occurs with certain verbs which are specified by "L:verbs.txt" ('verbs.txt' is a list of verbs). For the above Hindi sentence post-position 'ko' can perhaps be introduced, i.e. 'vaha dilli ***ko*** gayaa', but 'vaha dilli gayaa' is more natural, and the translated sentence is better off without a 'ko'.

[2] List

Finally, each preposition handled has a default rule, which is applied at the end when all the other rules for that preposition fail; the sense given by the default rule is based on the most frequent usage of the preposition at the target side. All the fields (except the first and last) in the default rule have hyphens. The default rule for *'to'* is written below;

to, -, -, -, -, to%p%1

Some of the rules in the rule file are given below, for ease of comprehension, we mention the actual target sense instead of the dictionary id for the last field (the actual rule file has dictionary sense id)

at, v, L:peer_verbs.txt, n, -, kii tarapha
at, v, L:transaction_verbs.txt, n, price, meM
for, v, -, n, distance, taka
in, n, animate, n, place, kaa
on, v, -, n, time, ko
to, v, L:go_verbs.txt, n, animate|authority, ke paasa
with, v, -, n, instrument, se

2.3 Recap

We briefly describe the various steps of the algorithm again;

(a) Given a raw sentence we feed it to the Shakti MT system which performs various source language analysis, for our algorithm, information such as PP attachment and correct identification of the phrasal verb (if present) is crucial.
(b) The output of step (a) is taken by our module which automatically constructs the six field tuple described above. At this point we can only fill some fields, which are field 1 (source language preposition), field 2 (modified category) and field 4 (modifier category).
(c) We then compare this constructed tuple with the appropriate tuples present in the rule file. For this constructed tuple to satisfy the various constraints mentioned in the tuple with which it is compared resources such as place filter, time filter, lists and WordNet are consulted automati-

cally. The order in which we use these resources has been already been mentioned in section 2.1. The tuple for which all the constraints are satisfied is selected, the last field of this tuple contains the dictionary id of the sense.

(d) Output the selected sense.

3 Evaluation

For the current study, experiments were conducted with 6 high frequency prepositions, they are; *at*, *for*, *in*, *on*, *to*, and *with*. The algorithm was tested on 100 sentences for each preposition in both the language pairs, i.e., 600 sentences for English-Hindi and 600 sentences for English-Telugu. These sentences were randomly extracted from the ERDC[3] corpus. The corpus contains text from different domains such as medicine, sports, history, etc. The input to the implemented system was manually checked and corrected to make sure that there were no errors in the information which is expected by the system. The bulk of these corrections involved rectifying the wrong PP attachment given by the parser and the mistakes in phrasal verb identification.

Prep[4]	**Precision**	**BL**	**No. of Sense**
At	73.4	51.5	5
For	84.05	69.5	6
In	82	65.2	7
On	85	70	3
To	65.2	35.4	10
With	66	50	6

Table 1{English-Hindi}.

Prep[4]	**Precision**	**BL**	**No. of Sense**
At	68	48	5
For	72	50	7
In	82	82	3
On	76	76	2
To	80	80	2
With	94	90	3

Table 2{English-Telugu}.

3.1 Performance

The tables above show the performance of the system and compares it with the baseline score (BL). BL is the precision of the system with only the default sense. The tables also show the number of sense which English prepositions can take on the target side. Table 1 and Table 2 show English-Hindi and English-Telugu results respectively.

The implemented system gives very promising results. Certain prepositions give comparably low precision. The reasons for the inappropriate sense selection are discussed in the next section. The English-Telugu results (Table 2) show same system precision and BL for some preposition ('in' and 'to'). This is because these prepositions have less number of sense on the target side and all the instances found in the test data had the default sense.

3.2 Error analysis

The errors made by the system were analyzed and the major reasons for inappropriate sense selection were;

(a) Noise generated by WordNet,
(b) Special constructions,
(c) Metonymy,
(d) Ambiguous sentences,
(e) Presence of very general constraints.

The problem of noise generation by WordNet sometimes leads to surprising and unexpected sense selection; this is because in WordNet a noun or verb will have multiple sense, and each of these senses will have various levels of hypernym synsets, so, while finding various concepts/features (specified by the rule for a preposition) we need to look at each one of these senses. We need to do this because we currently don't have the sense information. So, an inappropriate sense might sometimes satisfy the constraint(s) and result in inappropriate selection. The solution for this will obviously be to identify the correct sense of modifier/modified prior to getting its semantic property from the WordNet.

There are certain constructions in which the head noun of the PP is a pronoun, which refers back to a noun. For us this will create a problem, in such cases we will first need to get the referent

[3] Electronic Research and Development Centre, NOIDA
[4] Prepositions

noun and then apply the constraints on it, take the following example;

(10) *The rate at which these reactions occur is known as rate of metabolism.*

In the above example, the head noun of the PP (*at which*) refers to the noun (*rate*) on which we need to apply the constraints. At present the coreference information is not available to us, therefore in such cases the algorithm fails to give the correct output.

The other reason for failure was the ambiguity of the sentence itself which could be interpreted in various ways, like the example below;

(11) *Andamaan should go to India.*

The above sentence can be interpreted (and translated) in two ways, the hindi translations for the two interpretation are;

(11a) '*andamaan indiaa ko jaanaa chahiye'*
'Andamaan' 'India' 'to' 'go' 'should'
India should get Andaman.

(11b) '*andamaan ko indiaa jaanaa chahiye'*
'Andamaan' 'to' 'India' 'visit' 'should'
Andaman should visit India.

In (11a) we get the sense that the possesion/control of 'Andamaan' should go to 'India', and in (11b) it is 'Andamaan' (the government of 'Andamaan') which is going to 'India' (the government of India), as in, *The United States should go to UK*, also in (11b) we can have 'Andamaan' as somebodys' name, as in, *Ram should go to India.* In such cases we failed to get the appropriate translation of the preposition as it in turn depends on the correct interpretation of the whole sentence. Ambiguity of numerals in a sentence is yet another case which lead to faliure, like the following example;

(12) *At 83, Vajpayee is overweight.*

In the above sentence, the number 83 can either mean this persons' (*Vajpayee*) age or his weight. The target side translation takes different preposition sense for these two interpretation. Hindi takes *para* and *in* Telugu 'at' is not-translated when we treat 83 as weight, and when treated as age, we get *mem* and *lo/ki* in Hindi and Telugu respectively.

We found that certain prepositions occur in large number of metonymical usage, like, 'with' and 'at'. The constraints in a rule have been formulated for the general usage and not the extended usage of a given word. The example below shows one such instance;

(13) *Great bowlers spend hours after hours at the nets.*

While looking in WordNet for the various senses of 'net' not a single sense matches with the kind of usage in which 'net' is used in the above sentence.

Certain rules for some of the preposition were found to be very general, the low performance of 'for' and 'to' in telugu and hindi respectively are mainly due to this reason. In general, formulating rules (English-Hindi) for preposition 'to' was very difficult. This was because 'to' can have around 10 senses in Hindi. The rules with very general constraints tend to satisfy cases where they should have failed. One has to revisit them and revise them.

4 Conclusion and Future Work

In this paper we described an approach to select the appropriate sense for a preposition from an English to Indian language MT perspective, we discussed the issues involved in the task, we explained the steps to achieve the required task; which are, semantic and context extraction, and sense selection. We reported the performance of the system, and showed that our approach gives promising results. We also discussed the identified problems during the error analysis; such as noise generation by WordNet.

One of the pertinent tasks for the future would be to come up with a solution to reduce the noise generated by WordNet. The scope of rule file in terms of handling more prepositions needs to be broadened. We would like to extend this work to handle complex preposition. Finally, we would like to explore if ML techniques can be combined with the rule base to exploit the benefits of both the approaches.

References

Yukiko Sasaki Alam. 2004. Decision Trees for Sense Disambiguation of Prepositions: Case of Over. In *HLT/NAACL-04.*

Marija M. Brala. 2000. Understanding and translating (spatial) prepositions: an exercise in cognitive semantics for lexicographic purposes.

Joseph Emonds. 1985. *A unified theory of syntactic categories.* Dordrecht: Foris.

Gilles Fauconnier. 1994. *Mental spaces*. Cambridge: Cambridge University Press.

M. Grimaud. 1988. Toponyms, Prepositions, and Cognative Maps in English and French, *Journal of American Society of Geolinguistics.*

Ebba Gustavii. 2005. Target Language Preposition Selection - an Experiment with Transformation-based Learning and Aligned Bilingual Data. In *Proceedings of EAMT 2005*.

Sven Hartrumpf, Hermann Helbig and Rainer Osswald. 2005. Semantic Interpretation of Prepositions for NLP Applications. In *EACL 2006 Workshop: Third ACL-SIGSEM Workshop on Prepositions*

A. Herskovits. 1986. *Language and Spatial Cognition, An Interdisciplinary Study of Prepositions in English*. Cambridge University Press.

Cliffort Hill. 1982. Up/down, front/back, left/right. A contrastive study of Hausa and English. In *Weissen born and Klein,* 13-42.

Ray Jackendoff. 1977. *The architecture of the language*. Cambridge, MA: MIT Press.

Nathalie Japkowicz and Janyce M. Wiebe. 1991. A System For Translating Locative Preposition From English Into French. In *Proc. Association for Computational Linguistics.*

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Beth Levin. 1993. *English verb classes and alternations*. Chicago/London: The University of Chicago-Press.

George A. Miller. 1990. WordNet: An online lexical database. *International Journal of Lexicography.*

Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2005. Handling of Prepositions in English to Bengali Machine Translation. In *EACL 2006 Workshop: Third ACL-SIGSEM Workshop on Prepositions*.

Geoffrey Pullum and Rodney Huddleston. 2002. Prepositions and prepositional phrases. In *Huddleston and Pullum (eds.),* 597-661.

Gisa Rauh. 1993. On the grammar of lexical and nonlexical prepositions in English. In *Zelinskiy-Wibbelt (eds.)*, 99-150.

Patrick Saint-Dizier and Gloria Vazquez. 2001. A compositional framework for prepositions. In *IWCS4, Tilburg, Springer*, lecture notes, p. 165-179.

Patrick Saint-Dizier. 2005. PrepNet: A framework for describing prepositions: Preliminary investigation results. In *Proc. of IWCS 6*, Tilburg.

Joseph M. Sopena, Agusti LLoberas and Joan L. Moliner. 1998. A connectionist approach to prepositional phrase attachment for real world texts. In *COLING-ACL '98*, 1233-1237.

T. Tezuka, R. Lee, H. Takakura, and Y. Kambayashi. 2001. Web-Based Inference Rules for Processing Conceptual Geo-graphical Relationships. In *Proc. of the 2nd Int. Conf. on Web Information Systems Engineering, The 1st Int. Workshop on Web Geographical Information Systems*.

Arturo Trujillo. 1992. Locations in the machine translation of prepositional phrases. In *Proc. TMI-92.*