Translating Lexical Semantic Relations:
The First Step Towards Multilingual Wordnets[*]

Chu-Ren Huang, I-Ju E. Tseng, Dylan B.S. Tsai
Institute of Linguistics, Preparatory Office, Academia Sinica
128 Sec.2 Academy Rd., Nangkang, Taipei, 115, Taiwan, R.O.C.
churen@gate.sinica.edu.tw, {elanna, dylan}@hp.iis.sinica.edu.tw

Abstract

Establishing correspondences between wordnets of different languages is essential to both multilingual knowledge processing and for bootstrapping wordnets of low-density languages. We claim that such correspondences must be based on lexical semantic relations, rather than top ontology or word translations. In particular, we define a translation equivalence relation as a bilingual lexical semantic relation. Such relations can then be part of a logical entailment predicting whether source language semantic relations will hold in a target language or not. Our claim is tested with a study of 210 Chinese lexical lemmas and their possible semantic relations links bootstrapped from the Princeton WordNet. The results show that lexical semantic relation translations are indeed highly precise when they are logically inferable.

1. Introduction

A semantic network is critical to knowledge processing, including all NLP and Semantic Web applications. The construction of semantic networks, however, is notoriously difficult for ‘small’ (or ‘low-density’) languages. For these languages, the poverty of language resources, and the lack of prospect of material gains for infrastructure work conspire to create a vicious circle. This means that the construction of a semantic network for any small language must start from scratch and faces inhibitive financial and linguistic challenges.

In addition, semantic networks serve as reliable ontolog(ies) for knowledge processing only if their conceptual bases are valid and logically inferable across different languages. Take wordnets (Fellbaum 1998), the *de facto* standard for linguistic ontology, for example. Wordnets express ontology via a network of words linked by lexical semantic relations. Since these words are by definition the lexicon of each language, the wordnet design feature ensures versatility in faithfully and comprehensively representing the semantic content of each language. Hence, on one hand, these conceptual atoms reflect linguistic idiosyncrasies; on the other hand, the lexical semantic relations (LSR’s) receive universal interpretation across different languages. For example, the definition of relations such as synonymy or hypernymy is universal. The universality of the LSR’s is the foundation that allows wordnet to serve as a potential common semantic network representation for all languages. The premise is tacit in Princeton WordNet (WN), EuroWordNet (EWN, Vossen 1998), and MultiWordNet (MWN, Pianta et al. 2002). It is also spelled out explicitly in the adaptation of LSR tests for Chinese (Huang et al. 2001).

Given that LSR’s are semantic primitives applicable to all language wordnets, and that the solution to the low-density problem in building language wordnets must involve bootstrapping from another language, LSR’s seem to be the natural units for such bootstrapping operations. The rich and structured semantic information described in WN and EWN can be transported through accurate translation if the conceptual relations defined by LSRs remain constant in both languages. In practice, such an application would also serve the dual purpose of creating a bilingual wordnet in the process.

In this paper, we will examine the validity of cross-lingual LSR inferences by bootstrapping a Chinese Wordnet with WN. In practice, this small-scale experiment shows how a wordnet for a low-density language can be built through

[*] An earlier version of this paper was presented at the Third Chinese Lexical Semantics Workshop at Academia Sinica in May 2002. We are indebted to the participants as well as colleagues at CKIP for their comments. We would also like to thank the SemaNet 2002 reviewers for their helpful comments. It is our own responsibilities that, due to the short revision time, we were not able to incorporate all their suggestions, especially comparative studies with some relative GWA papers. We are also responsible for all remaining errors

bootstrapping from an available wordnet. In theoretical terms, we explore the logical conditions for the cross-lingual inference of LSR's.

2. Translation Equivalents and Semantic Relations

Note that two translation equivalents (TE) in a pair of languages stand in a lexical semantic relation. The most desirable scenario is that when the two TE's are synonymous, such as the English 'apple' to the Mandarin 'ping2guo3'. However, since the conceptual space is not segmented identically for all languages, TE's may often stand in other relations to each other. For instance, the Mandarin 'zuo1zhi5' is a hypernym for both the English 'desk' and 'table'. Suppose we postulate that the LSR's between TE's are exactly identical in nature to the monolingual LSR's described in wordnets. This means that the lexical semantic relation introduced by translation can be combined with monolingual LRS's. Predicting LSR's in a target language based on source language data become a simple logical operation of combining relational functions when the LSR of translation equivalency is defined. This framework is illustrated in Diagram 1.

$$\begin{tikzcd}
CW_2 & EW_2 \arrow[l, "ii"'] \\
CW_1 \arrow[u, dashed, "y"] \arrow[r, "i"'] & EW_1 \arrow[u, "x"']
\end{tikzcd}$$

$x = EW_1 - EW_2$	LSR
$y = CW_1 - CW_2$	LSR
$i = CW_1 - EW_1$	Translation LSR
$ii = CW_2 - EW_2$	Translation LSR

The unknown LSR $y = i + x + ii$

Diagram 1. Translation-mediated LSR Prediction (The complete model)

CW_1 represents our starting Chinese lemma which can be linked to EW_1 through the translation LSR *i*. The linked EW_1 can than provide a set of LSR predictions based on the English WN. Assume that we take the LSR *x*, which is linked to EW_2. That LSR prediction is mapped back to Chinese when EW_2 is translated to CW_2 with a translation LSR *ii*. In this model, the relation *y*, between CW_1 and CW_2 is a functional combination of the three LSR's *i*, *x*, and *ii*.

However, it is well known that language translation involves more than semantic correspondences. Social and cultural factors also play a role in (human) choices of translation equivalents. It is not the aim of this paper to predict when or how these semantically non-identical translations arise. The aim is to see how much lexical semantic information is inferable across different languages, regardless of translational idiosyncrasies. In this model, the prediction relies crucially on the semantic information provided by the source language (e.g. English) lexical entry as well as the lexical semantic correspondence of a target language (e.g. Chinese) entry. The translation relations of the relational target pairs, although capable of introducing more idiosyncrasies, are not directly involved in the prediction. Hence we make the generalization that any discrepancy introduced at this level does not affect the logical relation of LSR prediction and adopt a working model described in Diagram 2. We only take into consideration those cases where the translation LSR ii is exactly equivalent, i.e., $EW_2 = CW_2$. This step also allows us to reduce the maximal number of LSR combination in each prediction to two. Thus we are able to better predict the contribution of each mono- or bi-lingual LSR.

$$\begin{tikzcd}
 & EW_2 = CW_2\ (ii = 0) \\
\circ \arrow[ur, dashed, "y"] \arrow[r, "i"'] & EW_1 \arrow[u, "x"']
\end{tikzcd}$$

The unknown LSR $y = i + x$

Diagram 2. Translation-mediated LSR Prediction (Reduced Model, currently adopted)

2.1 LRS Inference as Relational Combination

With the semantic contribution of the translation equivalency defined as a (bilingual) LSR, the inference of LSR in the target language wordnet is a simple combination of semantic relations. The default and ideal situation is where the two TE's are synonymous.

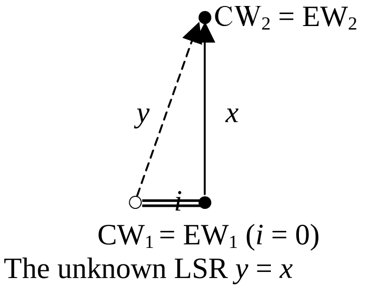


Diagram 3. Translation-mediated LSR Prediction (when TE's are synonymous)

In this case, the translation LSR is an identical relation; the LSR of the source language wordnet can be directly inherited. This is illustrated in Diagram 3.

However, when the translation has a non-identical semantic relation, such as antonyms and hypernyms, then the LSR predicted is the combination of the bilingual relation and the monolingual relation. In this paper, we will concentrate on Hypernyms and Hyponyms. The choice is made because these two LSR's are transitive relations by definition and allows clear logical predications when combined. The same, with some qualifications, may apply to the Holonym relations. Combinations of other LSR's may not yield clear logical entailments. The scenarios involving Hyponymy and Hypernymy will be discussed in section 3.3.

3. Cross-lingual LSR Inference: A Study based on English-Chinese Correspondences

In this study, we start with a WN-based English-Chinese Translation Equivalents Database (TEDB)[1]. Each translation equivalents pair was based on a WN synset. For quality control, we mark each TE pair for its accuracy as well as the translation semantic relation.

For this study, the 200 most frequently used Chinese words plus 10 adjectives are chosen (since there is no adjective among the top 200 words in Mandarin). Among the 210 input lemmas, 179 lemmas[2] find translation equivalents in the TEDB and are mapped to 497 English synsets. The occurring distribution is as follows: 84 N's with 195 times; 41 V's with 161 times; 10 Adj's with 47 times; and 47 Adv's with 94 times. 441 distinct English synsets are covered under this process, since some of the TE's are for the same synset. This means that each input Chinese lemma linked to 2.4 English synsets in average. Based on the TEDB and English WN, the 179 mapped input Chinese lemmas expanded to 597 synonyms. And extending from the 441 English synsets, there are 1056 semantically related synsets in WN, which yields 1743 Chinese words with our TEDB.

3.1. Evaluation of the Semantics of Translation

Six evaluative tags are assigned for the TEDB. Four of them are remarks for future processing. The LSR marked are

- **Synonymous**: TE's that are semantically equivalent.
- **Other Relation**: TE's that hold other semantic relations

The result of evaluation of TE's involving the 210 chosen lemma are given in Table 1.

	Syn.	Incorrect	Other Relation	Total
N	148	32	15	195
	75.90%	16.41%	7.69%	100%
V	113	29	19	161
	70.18%	18.01%	11.8%	100%
Adj	39	8	0	47
	82.98%	17.02%	0%	100%
Adv	83	8	3	94
	88.3%	8.51%	3.19%	100%
Total	382	78	36	496
	77.02%	15.73%	7.26%	100%

Table 1. Input Lemmas (Total subject =496)

Illustrative examples of our evaluation are given below:

1a) Synonymous: *企業* ***qi4ye4*** (N) // ***enterprise:*** an organization created for business ventures
1b) Incorrect: *表示* ***biao3shi4*** (V) // '***extend***', '***offer***': make available; provide
1c) Other Relation: *市場* ***shi4chang3*** (N) //

[1] The translation equivalence database was hand-crafted by the CKIP WordNet team. For each of the 99642 English synset head words, three appropriate translation equivalents were chosen whenever possible. At the time when this study was carried out, 42606 TE's were proofed and available

[2] The input lemmas for which TE's were unable to find are demonstratives or pronouns for nouns, and aspect markers for adverbs

‘***market, securities_industry***’: the securities markets in the aggregate

Table 2 indicates the relations between the synonyms of an input lemma and the same English synset. Recall that our TEDB gives more than one Chinese translation equivalent to one English WN entry. Hence we can hypothesize that the set of Chinese translation equivalents form a synset. It is natural, then, to examine the semantic relations between other synset members and the original WN entry. Table 1 and 2 show a rather marked difference in terms of the correctness of the synonymy relation. This will be further explained later.

	Syn.	Incor.	Other Rel.	Others	Total
N	114	51	25	19	209
	54.5%	24.4%	11.0%	9.1%	100%
V	104	46	18	14	182
	57.1%	25.3%	9.99%	7.7%	100%
Adj	37	8	2	10	57
	64.9%	14.0%	3.5%	17.5%	100%
Adv	119	20	4	6	149
	79.9%	13.4%	2.7%	4.0%	100%
Total	374	125	49	49	597
	62.6%	20.9%	8.2%	8.2%	100%

Table 2. Synonyms of Input Lemma (Total Subject=597)

From the data above, we observe two generalizations: First, polysemous lemmas have lower possibility of being synonymous to the corresponding English synset. In addition, we also observe that there is a tendency for some groups, i.e., groups with polysemy and with abstract meanings, to match synonymous English synsets. These findings are helpful in our further studies when constructing CWN, as well as in the application of TEDB.

3.2 Cross-lingual LSR predictions with synonymous translations

The next step is to take the set of English LSR’s stipulated on a WN synset and transport them to its Chinese translation equivalents. We evaluated the validity of the inferred semantic relations in Chinese. In this study, we concentrated on three better-defined (and more frequently used) semantic relations: antonyms (ANT); hypernyms (HYP); and hyponyms (HPO). Here, we limit our examination to the Chinese lemmas that are both translation equivalents of an English WN entry and are considered to have synonymous semantic relations to that entry. The nominal and verbal statistics are given in Table 3 and Table 4 respectively.

	Syn.	Incor.	Other Rel.	Others	Total
ANT	7	3	0	2	12
	58.3%	25%	0%	16.7%	100%
HYP	117	33	15	20	185
	63.2%	17.8%	8.1%	10.8%	100%
HPO	284	119	66	256	725
	39.2%	16.4%	9.1%	35.3%	100%
Total	408	155	81	278	922
	44.3%	16.8%	8.8%	30.2%	100%

Table 3. Nouns (Total Number of Inferable Semantic Relations=922)

	Syn.	Incor.	Other Rel.	Others	Total
ANT	8	6	0	9	23
	34.8%	26.1%	0%	39.1%	100%
HYP	61	18	6	2	87
	70.1%	20.7%	6.9%	2.3%	100%
HPO	118	81	19	74	292
	40.4%	27.7%	6.5%	25.3%	100%
Total	187	105	25	85	402
	46.5%	26.1%	6.2%	21.1%	100%

Table 4. Verbs (Total Number of Inferable Semantic Relations=402)

From the 148 nouns where the English and Chinese translation equivalents are also synonymous, there are 357 pairs of semantic relations that are marked in English WN and are therefore candidates for inferred relations in Chinese. On average, each nominal RC translation equivalent yields 2.41 inferable semantic relations. The precision of the inferred semantic relation is tabulated below.

	Correct		Others		Total	
ANT	8	100%	0	0%	8	100%
HYP	70	79.5%	18	20.5%	88	100%

HPO	238	91.2%	23	8.8%	261	100%
Total	316	88.5%	41	11.5%	357	100%

Table 5. Precision of English-to-Chinese SR Inference (Nouns)

The study here shows that when no additional relational distance is introduced by translation (i.e. the 75.9% of nominal cases when TE's are synonyms), up to 90% precision can be achieved for bilingual LSR inference. And among the semantic relations examined, antonymous relations are the most reliable when transportabled cross-linguistically.

For the 112 verbs where the English and Chinese TE's are synonymous, there are 155 pairs of semantic relations that are marked in WN and are therefore candidates for inferred relations in Chinese. In contrast to nominal translation equivalents, each pair of verbal TE only yields 1.38 inferable semantic relations. The precision of the inferred semantic relation is tabulated in Table 6.

	Correct		Incorrect		Total	
ANT	14	100%	0	0%	14	100%
HYP	35	70%	15	30%	50	100%
HPO	75	82.4%	16	17.6%	91	100%
Total	124	80%	31	20%	155	100%

Table 6. Precision of English-to-Chinese SR Inference (Verbs)

Similar to the results of nouns, antonymous relations appear reliable in the behaviors of verbs as well. As to the other types of relations, the correct rates seem to be slightly lower than nouns. The precision for English-to-Chinese semantic relation inference is 80% for verbs.

The observed discrepancy in terms of semantic relations inference between nouns and verbs deserves in-depth examination. Firstly, the precision of nominal inference is 8.52% higher than verbal inference. Secondly, the contrast may not be attributed to a specific semantic relation. Both nouns and verbs have the same precision pattern for the three semantic relations that we studied. Inference of antonymous relations is highly reliable in both categories (both 100%). Hyponymous inference is second, and about 12% higher than hypernymous inference in each category (the difference is 11.64% for nouns and 12.42% for verbs). And, last but not least, the precision gaps between nouns and verbs, when applicable, are similar for different semantic relations (9.55% for hypernyms and 8.77% for hyponyms). All the above facts support the generalization that nominal semantic relations are more reliably inferred cross-linguistically than verbal semantic relations. A plausible explanation of this generalization is the difference in mutability of nominal and verbal meanings, as reported by Ahrens (1999). Ahrens demonstrated with off-line experiments that verb meanings are more mutable than noun meanings. She also reported that verb meanings have the tendency to change under coercive contexts. We may assume that making the cross-lingual transfer is a coercive context in terms of meaning identification. Taking the mutability into account, we can predict that since verb meanings are more likely than nouns to change under given coercive conditions, the changes will affect their semantic relations. Hence the precision for semantic relations inference is lower for verbs than for nouns.

In the above discussion, we observed that the three semantic relations seem to offer clear generalizations with regard to the precision of the inferences, as shown in Table 7.

	Correct		Incorrect		Total	
ANT	22	100%	0	0%	22	100%
HYP	105	76.1%	33	13.9%	138	100%
HPO	313	88.9%	39	11.1%	352	100%
Total	440	85.9%	72	14.1%	512	100%

Table 7. Combined Precision of English-to-Chinese SR Inference (Nouns+Verbs)

Two generalizations emerge from the above data and call for explanation: First, inference of antonymous relations is highly reliable; second, inference of hypernymous relations is more reliable than inference of hyponymous relations.

The fact that inference of antonymous relations is highly precise may be due to either of the following facts. Since the number of antonymic relations encoded is relatively few (only 22 all together), they may all be the most prototypical case. In addition, a pair of antonyms by definition differs in only one semantic feature and has the shortest semantic distance between them. In other words, an antonym (of any word) is simply a privileged (near) synonym whose meaning offers contrast at one particular semantic dimension. Since antonymy presupposes synonymous relations, it preserves the premise of our current semantic relation inference.

The fact that hyponymous relations can be more reliably inferred cross-linguistically than hypernymous relations is somewhat surprising, since they are symmetric semantic relations. That is, if A is a hypernym of B, then B is a hyponym of A. Logically, there does not seem to be any reason for the two relations to have disjoint distributions when transported to another language. However, more careful study of the conceptual nature of the semantic relations yields a plausible explanation.

We should take note of the two following facts: First, a hyponym link defined on an English word Y presupposes a conceptual class denoted by Y, and stipulates that Z is a kind of Y (see Diagram 4).

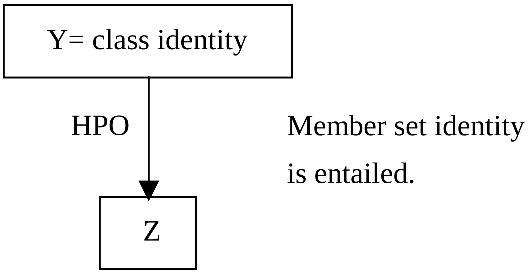


Diagram 4. class vs. member identity (HPO)

Second, a hypernym link defined on Y presupposes an identity class X which is NOT explicitly denoted, and stipulates that Y is a kind of X (see Diagram 5). Hence, it is possible that there is another valid conceptual class W in the target language that Y is a member of. And yet W is not equivalent to X.

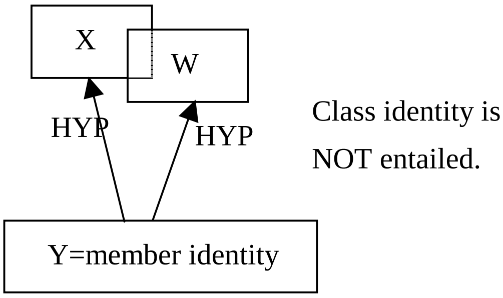


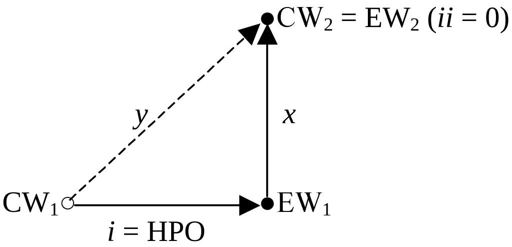
Diagram 5. class vs. member identity (HYP)

Since our inference is based on the synonymous relation of the Chinese TE to the English word Y, the conceptual foundation of the semantic relation is largely preserved, and the inference has a high precision. The failure of inference can in most cases be attributed to the fact that the intended HYP has no synonymous TE in Chinese. To infer a hyponymous relation, however, we need to presuppose the trans-lingual equivalence of the conceptual class defined by HPO. And since our inference only presupposes the synonymous relation of Y and its TE, and says nothing about HPO, the success of inference of the hyponymous relation is than dependent upon an additional semantic condition. Hence that it will have lower precision can be expected.

To sum up, our preliminary evaluation found that the precision of cross-lingual inference of semantic relation can be higher than 90% if the inference does not require other conceptual/semantic relations other than the synonymy of the translation equivalents. On the other hand, an additional semantic relation, such as the equivalence of the hypernym node in both languages when inferring hyponym relations, seems to bring down the precision rate by about 10%.

3.3. When Translation Introduces an additional LSR

In this section, we study the cases where translation introduces a hypernymous/ hyponymous LSR. These cases offer the real test to our proposal that TE's be treated as bilingual LSR's. The LSR inference here refuses non-vacuous combinations of two LSR's. For 37 Chinese input lemmas that hold other relations with English synsets, 57 semantically related links were expanded. First, we investigated the situation when the English synset occurs as a hyponym of the Chinese input lemma (Diagram 6).



(a) IF *x* = HPO
y = HPO + HPO = HPO (Hyponym is transitive.)

(b) IF *x* = HYP
y = HPO + HYP = ?

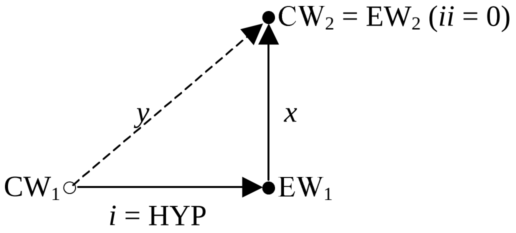
Diagram 6. Predicting LSR, when English is the hyponym of Chinese translation

33 inferable relations satisfied above description. Among them, 8 falls in the entailment of figure

6(a). Manual evaluation confirms the prediction. The other 25 cases are not logically inferable and do indeed show a range of different relations. The logically entailed HPO relation is exemplified below:

- *吃* ***chi1*** HPO→ [eat, feed] take in food HPO→ [raven] feed greedily TE→ *狼吞虎嚥* ***lang2tuen1hu3yan4***
  SO, *吃* ***chi1*** HPO→ *狼吞虎嚥* ***lang2 tuen1hu3yan4***

Next, when an English synset is marked as a hypernym to the Chinese input lemma, logically, hypernymous relation is transitive (Diagram 7).



(a) IF x = HYP
NOTE: y = HYP + HYP = HYP
(Hypernym is transitive.)
(b) IF x = HPO
NOTE: y = HYP + HPO = ?

Diagram 7. Predicting LSR, when English is the hypernym of Chinese translation

We found 2 cases (actually expanded from the same synset) under this condition.

- *使* ***shi3*** HYP→ [leave] cause to be in a specified state HYP→ [get, make] give certain properties to something TE→ *使* ***shi3*** / *致使* ***zhi4shi3***

Note that the same Chinese word 使 shi3 is used for both the head word and its hypernym. Hence, there are two possible interpretations of the data. The first possibility is that Chinese simply has a coarser-grain sense distinction in this case and the hypernym relation is incorrect. The second possibility is that the relation is self-hypernym (Fellbaum 1999). Since a fine-grain sense distinction is beyond the scope of the current paper, we will not decide on either interpretation.

In sum, our lexical semantic relation model makes correct distinctions among inferable and non-inferable LSR's. More specifically, it has a 100% prediction for hyponymous relations. For hypernymous relations even though the logical entailment could not be verified due to sparseness of data; it did correctly predict the portion of data that was logically non-inferable. We expect future studies with a wider set of data to confirm this prediction.

4. Conclusion

In this paper, we proposed to treat the translation equivalents relations as a set of bilingual lexical semantic relations. This proposal allows us to process bi-lingual inference of LSR's as simple functional combinations of semantic relations. The process itself greatly reduces the complexity of bootstrapping wordnets from a different language. We empirically supported our proposal by successfully applying it to the inference of Chinese LSR's from English WN.

The proposed approach requires bilingual TEDB's that are marked with translation semantic relations. Although such TEDB's are not widely available yet, they are necessary for cross-lingual language processing such as MT and IR, as well as for any type of knowledge processing. We hope that our approach can promote the construction of LSR-marked TEDB as well as multilingual wordnets.

References:

Ahrens K. 1999. *The Mutability of Noun and Verb Meaning*. Chinese Language and Linguistics V. Interactions in Language, Y. Yin, I. Yang, & H. Chan (eds.), pp. 335 – 548. Taipei. Academia Sinica.

Fellbaum, C. (ed.). 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Huang, Chu-Ren. 2000. *Towards a Chinese Wordnet and a CE/EC Bi-Wordnet*. Chinese Language Sciences Workshop: Lexical Semantics. October 9, 2000. Department of Chinese, Translation and Linguistics. City University of Hong Kong.

Huang, Chu-Ren, D. B. Tsai, J. Lin, S. Tseng, K.J. Chen, and Y. Chuang. 2001. *Definition and Test for Lexical Semantic Relations in Chinese*. [in Chinese] Paper presented at the Second Chinese Lexical Semantics Workshop. May 2001, Beijing, China.

Pianta, Emanuel, L. Benitivogli, C. Girardi. 2002 *MultiWordNet: Developing an aligned multilingual database*. Proceedings of the 1st International WordNet Conference, Mysore, India, pp. 293-302.

Vossen P. (ed.). 1998. EuroWordNet: A multilingual database with lexical semantic networks. Norwell, MA: Kluwer Academic Publishers.