Machine Translation **17:** 99–137, 2002.

Construction of a Chinese–English Verb Lexicon for Machine Translation and Embedded Multilingual Applications

BONNIE JEAN DORR and GINA-ANNE LEVOW
University of Maryland Institute for Advanced Computer Studies, College Park, MD 20742, U.S.A. (E-mails: bonnie@umiacs.umd.edu; gina@umiacs.umd.edu)

DEKANG LIN
University of Alberta, Edmonton, Alberta, Canada, T6G 2E8 (E-mail: lindek@cs.ualberta.edu)

Abstract. This paper addresses the problem of automatic acquisition of lexical knowledge for rapid construction of engines for machine translation and embedded multilingual applications. We describe new techniques for large-scale construction of a Chinese–English verb lexicon and we evaluate the coverage and effectiveness of the resulting lexicon. Leveraging off an existing Chinese conceptual database called HowNet and a large, semantically rich English verb database, we use thematic-role information to create links between Chinese concepts and English classes. We apply the metrics of recall and precision to evaluate the coverage and effectiveness of the linguistic resources. The results of this work indicate that: (a) we are able to obtain reliable Chinese–English entries both with and without pre-existing semantic links between the two languages; (b) if we have pre-existing semantic links, we are able to produce a more robust lexical resource by merging these with our semantically rich English database; (c) in our comparisons with manual lexicon creation, our automatic techniques were shown to achieve 62% precision, compared to a much lower precision of 10% for arbitrary assignment of semantic links.

Key words: resource alignment, Chinese–English lexicons, thematic roles, lexical acquisition

1. Introduction

The growing quantity of electronically available multilingual information has created an urgent need for rapid construction of semantic resources for language analysis. It is infeasible to produce large-scale repositories of semantic representations for multiple languages using human labor alone: construction of large semantic lexicons is a slow, tedious, and error-prone process (Viegas et al., 1996). As a result, many researchers have applied *semi-automatic* techniques to assign semantic classes to words (Dorr et al., 1997; Palmer and Wu, 1995; Palmer and Rosenzweig, 1996; Palmer et al., 1997) using a variety of online resources including *Longman's Dictionary of Contemporary English* (Procter, 1978), *English Verb Classes and Alternations* (Levin, 1993), and WordNet (Miller and Fellbaum, 1991; Fellbaum, 1998).

Since large monolingual resources have become increasingly accessible to the research community, the time is ripe for the development of *automatic* techniques for semantic lexicon construction. Several researchers have already investigated resource alignment for this purpose (Palmer and Wu, 1995; Peters et al., 1998; Vossen et al., 1997 and Carpuat et al., 2002). However, these approaches are either small-scale in nature – or the broader-scale approaches do not automate the acquisition of argument-structure information (i.e., thematic roles) across languages.

This paper describes an alternative approach that exploits the availability of online semantic resources to produce a single foreign-language semantic lexicon with thematic-role information. We demonstrate that automatic linking of lexical entries from two different languages can be achieved through the use of a lexical-semantic classification which provides a framework for verb disambiguation based on thematic roles (Dorr, 2001). We view this work to be a significant enhancement to current efforts in resource building for multilingual applications such as EuroWordNet (Vossen, 1998; Vossen et al., 1998), PropBank (Palmer et al., 2001) and the Chinese Propbank under development (Palmer et al., 2002).

Our initial language pair is Chinese–English, as it is well known in the Human Language Technology (HLT) community that Chinese is expected to become the number one Web language by 2007. However, we also expect our techniques to be more broadly applicable to other language pairs. This will become increasingly important since other languages are also expected to surpass English in quantity: "Of the World's 6 Billion People, only 7% speak English".[1] We focus specifically on verbs (e.g., *transport*) – and their nominalized forms (e.g., *transportation*). These are the most complex units in multilingual processing: verbs are at the heart of a wide range of cross-language *divergences* that occur in as much as 35% of multilingual corpora (Dorr et al., 2002).

The semantic lexicon resulting from the application of our lexical-acquisition procedure significantly impacts the effectiveness of the translation of multiply ambiguous Chinese verbs. The importance of determining the appropriate word sense in machine translation (MT) is clear when one considers the degree of inaccuracy that might result from using a weak alternative, such as access to a bilingual word list.

As an example, the Chinese verb 拉 *la* has the following English translations: *help*, *transport*, *involve*, *implicate*, *pull*, *drag*, *chat*, *defecate*, *slash*, *cut*, *raise*, *move*, and *pressgang*.[2] Our semantic lexicon makes it possible to distinguish among these cases by assigning semantic roles from the lexicon as a part of the Chinese analysis component. For example, consider the Chinese sentences in (1) and (2):

(1) 张三在李四马上要输了的时候拉了他一把

Zhangsan zai Lisi ma-shang-yao shu le de shihou la le ta yi ba.

ZHANGSAN AT LISI ABOUT-TO LOSE, asp subord TIME PULL asp HIM ONE HAND

'When Lisi was just about to lose, Zhangsan helped him.'

(2) 张三拉一吨米到村子

Zhangsan la yi dun mi dao cunzi.

ZHANGSAN PULL ONE TON GRAIN TO VILLAGE

'Zhangsan transported a ton of grain to the village.'

The translations correspond to the lexical entries (3) and (4) in our automatically acquired Chinese semantic lexicon

(3) 拉 13.4.2 `(agent, theme, mod-poss)` help

(4) 拉 11.1 `(agent, theme, goal, source)` transport

where the thematic roles are filled in as in (5) and (6).

(5) (agent[Zhangsan],theme[him],mod-poss[When Lisi was about to lose])

(6) (agent[Zhangsan],theme[ton of grain],source,goal[to the village])

In the next section, we provide the background for our approach, including: (a) a description of the MT application for which it is used; (b) the use of this technology as an embedded component of a cross-language information retrieval system; and (c) a description of the representations and resources used to construct the Chinese semantic lexicon. After this, we describe the mapping between existing online resources.

Next, we demonstrate that our automatic acquisition techniques provide a framework for compensating for gaps in this new resource. We then evaluate the coverage and accuracy of the new Chinese lexicon with respect to the pre-existing resources, concluding that: (a) we are able to obtain reliable Chinese–English entries both with and without pre-existing semantic links between the two languages; (b) if we have pre-existing semantic links, we are able to produce a more robust lexical resource by merging these with our semantically rich English database.

After this, we discuss the issue of compensating for deficiencies in the existing resources and we apply the metrics of recall and precision – based on a human-judged gold-standard – to evaluate the coverage and effectiveness of our lexicon.

In our comparisons with manual lexicon creation, our automatic techniques were shown to achieve 62% precision (overall accuracy of the Chinese–English links we have already automatically acquired), compared to a much lower precision of 10% for arbitrary assignment of semantic links.

Finally, we relate our work to that of other researchers who have investigated the problem of mapping across semantic hierarchies for construction of MT resources.

2. Background

It was previously reported in Dorr (1998) that *manual* linking of lexical entries from two different languages – via WordNet senses – provides a significant ambiguity reduction in query translation for cross-language information retrieval. The work reported in this article builds on this basic idea, but provides an *automatic* technique for linking semantic concepts between two languages – using thematic roles instead of WordNet senses. We view the resulting system as an enabling technology for multilingual applications that include English as the user's language, i.e., the target language in MT or the query language in cross-language information retrieval.

Our approach is consistent with the generation-heavy philosophy of Habash and Dorr (2002), where English resources are rich enough to serve as the foundation for foreign-language applications. This approach has proven fruitful as the basis of embedded components for a variety of multilingual applications including cross-language information retrieval (Dorr and Katsova, 1998), foreign-language tutoring (Dorr, 1997a), and MT (Dorr et al., 1998).

Below, we describe the ChinMT (Chinese–English MT) system as a representative example of an application for which this enabling technology was developed. Following this, we outline the use of this technology in MADLIBS, an embedded component of a cross-language information retrieval system. Finally, we describe the LCS Verb Database (LVD) that serves as the foundation for the production of our Chinese semantic lexicon.

2.1. CHINESE–ENGLISH MT (CHINMT)

Our automatic acquisition routines were developed for construction of a Chinese semantic lexicon for the ChinMT system (Dorr et al., 1998; Olsen et al., 1998). ChinMT is an interlingual MT approach that uses thematic information to provide the surface realization of an English sentence. Automatic translation of Chinese documents requires the Chinese semantic lexicon in the analysis phase, after parsing for constructing the interlingual form of the source-language input sentence, and during lexical selection of generation of the target-language sentence(s).

Analysis in the ChinMT system relies on an in-house parser called REAP (Weinberg et al., 1995) to produce English parse trees on a large scale, extended to operate Chinese on a smaller scale. The parser output is semantically analyzed

using the Chinese semantic lexicon, producing a lexical representation that serves as the interlingua. Generation from this representation is achieved by means of a system called Oxygen (Habash, 2000), a variant of Nitrogen (Langkilde and Knight, 1998a,b,c) that combines our own linearizer implemented in Lisp with Nitrogen's statistical extraction module and Nitrogen's morphological generation engine.[3]

The English output is produced by means of two steps: lexical selection and syntactic realization. Lexical selection involves a comparison between components in the interlingua and abstract thematic roles associated with words in the English semantic lexicon. Syntactic realization recasts thematic roles as relations in an unordered tree where the root is an event concept and each child is linked by a relation. Generation of target-language sentences from the interlingua is described in more detail in Dorr et al. (1998).

A screen snapshot of a translation by ChinMT on a Chinese example is shown in Figure 1. This translation is more fluent than its literal (gisted) equivalent (7).

(7) Our Foreign_Economic_Trade_Ministry spokesperson lodge stern protest.

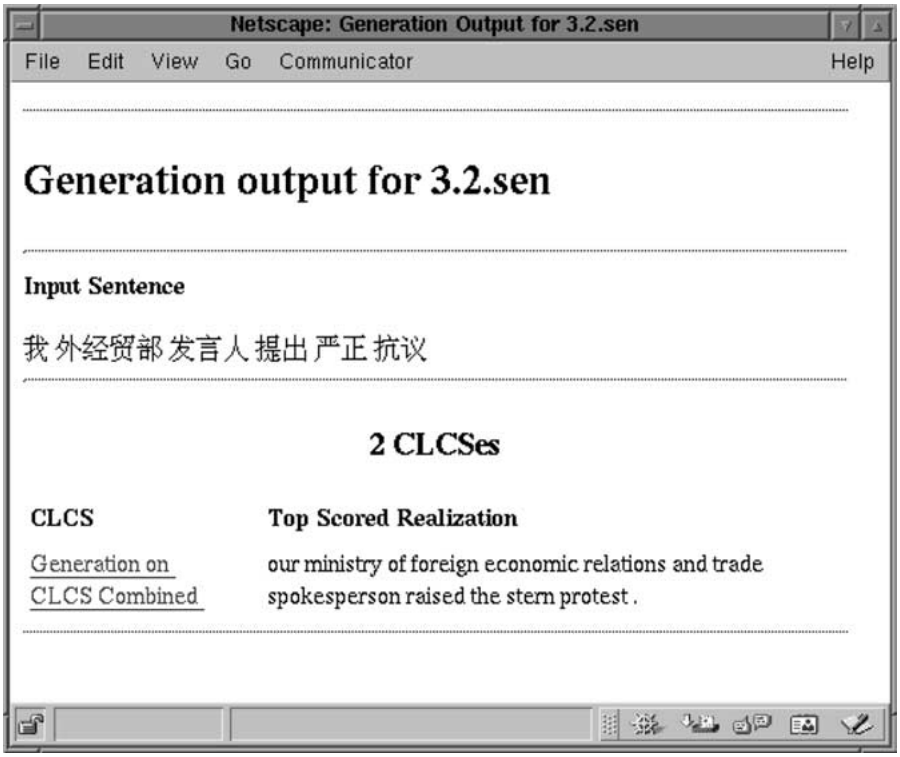

Figure 1. Translation of a Chinese sentence into English.

The automatic construction of the Chinese semantic lexicon – and small "clean-ups" – took a mere person-month of effort. In a recent evaluation, it was shown that the translation quality of this system is comparable to that of a commercial translation system that took seven years of effort to develop (Habash, 2003). Thus, the use of automatic techniques for development of semantic lexicons based on existing resources provides a significant time savings in the development of a new MT system (Habash and Dorr, 2001).

2.2. EMBEDDED MT: CROSS-LANGUAGE INFORMATION RETRIEVAL

Our Chinese semantic lexicon also serves as the basis for disambiguating and translating English terms into Chinese query terms for input to a retrieval system that allows the user to access Chinese documents using English as their query language. The importance of determining word senses in this embedded MT application is clear when one considers the degree of inaccuracy that might result from using a weak alternative, such as access to a bilingual word list. A variety of methods have been proposed to cope with this issue, also known as "translation ambiguity", where one source-language term translates into more than one target-language alternative (Oard and Dorr, 1996; Oard, 1998; Ballesteros and Croft, 1997; Hull and Grefenstette, 1996). These techniques include selecting every translation, the first n translations according to some ranking strategy, and those that co-occur with candidate translations of other terms in the query.

The technique we adopt – LCS Query Translation (LQT) – uses Lexical Conceptual Structure (LCS) from our semantic lexicon to transform the user's query into the document language for information retrieval. (The LCS is described below in Section 2.3.) We use a structured syntax interface, called MADLIBS (Maryland Action Detection: Language-Independent Browsing and Search), to ensure that the user's query is fully analyzable for application of the LQT technique. Specifically, for each word in the semantic lexicon, a set of simple "semantic templates" are pre-compiled based on the thematic-role frames (or "grids") associated with the word. At query time, the user specifies the event (i.e., verb) of interest and the semantic templates associated with that event are presented to the user in the form of a structured representation. The user then augments this representation by filling in argument positions with information relevant to the user's interest. The resulting augmented structure is passed to a structural-matching module of Oxygen (from the ChinMT system described above) and a Chinese bag of words is produced; this then serves as the query into the Chinese document collection.

The user interface for augmentation of the semantic template is illustrated in Figure 2. In this example, the user has selected the verb *provide* along with one of its three associated frames, "someone provided something with something". The positions corresponding to "someone" and "something" are associated with thematic roles for the verb *provide*; these three positions appear as empty boxes for free-form input from the user. In the current example, the user has typed *China*,

Taiwan, and *quake aid* as the argument fillers for the three roles. The interface allows querying of either English or Chinese documents; we will focus on the cross-language variant henceforth.

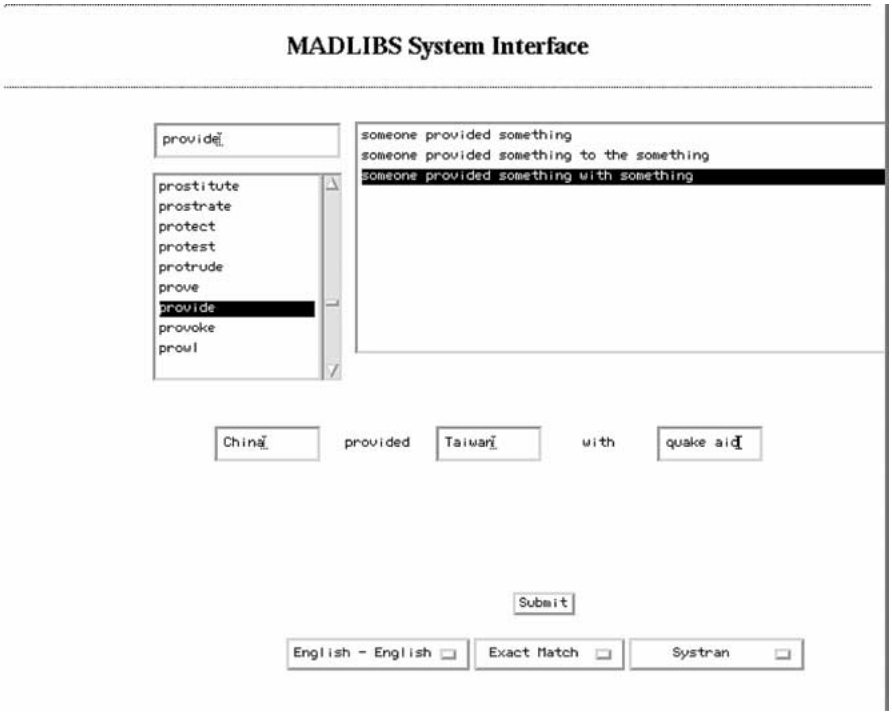

Figure 2. Structured syntax input interface.

The LCS (to be described next) is an important component of our LQT approach. In particular, construction of the foreign-language query involves generation from LCS-based templates which have been augmented by the user's input words. The correspondence between thematic structure and surface form constitutes an initial phase of sense disambiguation, identifying the subset of possible senses with this argument structure. To produce the final translation of the query, the Oxygen system applies structural matching of the query against a database of LCS structures in the foreign language. The translation terms identified by the structural match component of Oxygen form a bag of words that comprise the query to the retrieval system. We use a version of the Inquery 3.1p1 information retrieval system from the University of Massachusetts, modified for 2-byte encodings of Chinese characters.

Retrieval results are displayed interactively in a "Selection Interface". Documents may be displayed in the original document language or as a list of initial sentences translated by Systran (see Figure 3). The Selection Interface allows the user to choose one of the displayed documents for further inspection in a "Presentation Interface". The results are presented in a "gisted" (word-for-word translation)

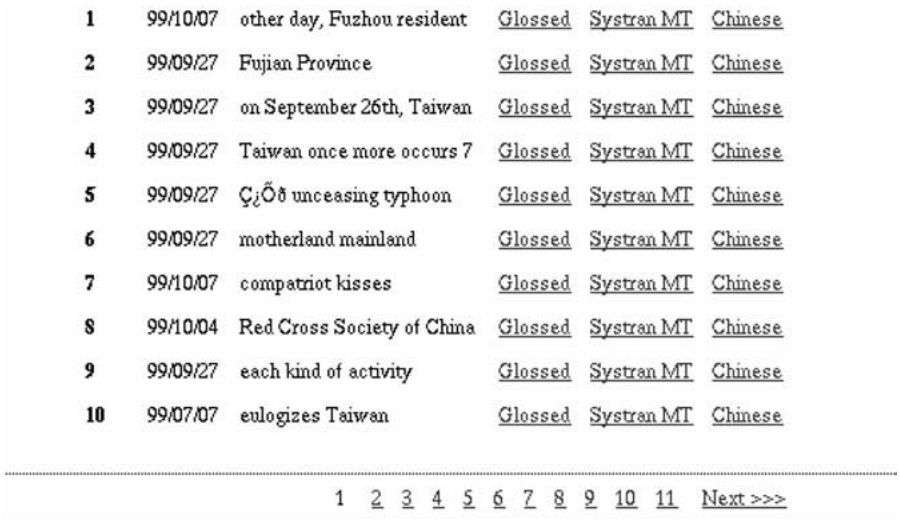

Figure 3. Selection interface.

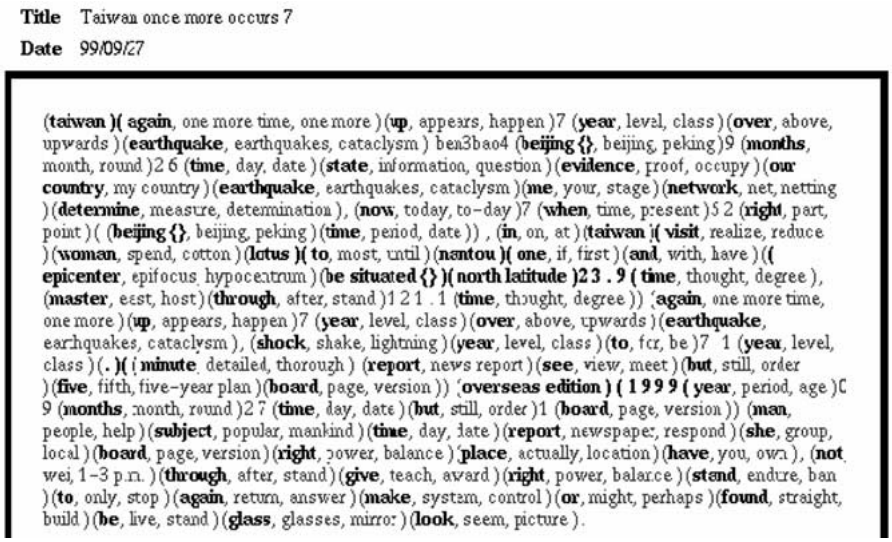

Figure 4. Presentation interface: "Gisted" format.

format where multiple ranked translations are displayed (see Figure 4). Additional details about the use of semantic representations in the CLIR system are given in Dorr and Katsova (1998) and Levow et al. (2000).

2.3. LCS VERB DATABASE (LVD)

Automatic construction of our Chinese semantic lexicon relies on three existing resources: (a) a classification of English verbs called the LCS Verb Database (LVD)

(Dorr, 2001); (b) a Chinese conceptual database called HowNet (Dong, 1988a,b,c); and (c) a large machine readable Chinese–English dictionary called Optilex. This subsection provides the background for the LVD; we defer the presentation of HowNet and Optilex for later sections.

Lexical-semantic knowledge is encoded in our semantic lexicon in the form of LCS – as formulated by Dorr (1993, 1994) based on work by Jackendoff (1983, 1990). This representation serves as the interlingua (language-independent structure) for the ChinMT and MADLIBS systems described above.

The LCS approach views semantic representation as a subset of conceptual structure, the language of mental representation. The representation includes "types" such as Event and State, which are specialized into "primitives" such as GO, STAY, BE, GO-EXT, and ORIENT. These are combined in different ways to provide the "semantic structure" for each class of verbs. The "semantic content" – typically represented as a manner component, e.g., [$_{\text{Manner}}$ TOUCH+INGLY]) – distinguishes among verbs within a semantic class, e.g., *touch*, *caress*, *nudge*, *pat*, and *stroke*. The full representation for (8a) is therefore the representation in (8b), roughly (8c). [4]

(8)a. John touched the cat.

b. [$_{\text{Event}}$ ACT_ONLoc
([$_{\text{Thing}}$ JOHN],
[$_{\text{Thing}}$ CAT],
[$_{\text{Manner}}$ TOUCH+INGLY])]

c. John acted on the cat by touching.

The LCS is the representation used in our Chinese and English semantic lexicons. The English semantic lexicon, the LVD, has now been publicly released for research purposes (Dorr, 2001). The verbs in this resource were borrowed initially from the online index for *English Verb Classes and Alternations* (*EVCA*) (Levin, 1993).[5] This index of verbs provides a shallow hierarchical structure of verb classes, where verbs that share syntactic behaviors are thus hypothesized to share a component of meaning, and thus are grouped together in the same class. For example, *steal* and *swipe* share the syntactic frames *X stole/swiped Y from Z* and *X stole/swiped Y for W*. These are grouped together in a class of verbs that share a component of meaning corresponding to "possessional deprivation".

While *EVCA* provides a unique and extensive catalog of verb classes, it does not define the underlying meaning components of each class. The LVD is the first resource to provide a relation between Levin's classes and meaning components, as defined by the LCS (Dorr, 1997b). Thus, the LVD is not a simple extension to *EVCA*, but an entirely new database with rich semantic structure.[6]

Beyond the LCS representation, the LVD includes additional information not encoded in *EVCA*, specifically: (a) 26 new classes (classes 000 through 026) for verb senses that were omitted from *EVCA* (Dorr, 1997a); (b) thematic roles for

each entry (Dorr et al., 1998); (c) WordNet senses for each *EVCA* verb (Dorr and Katsova, 1998); and (d) 3,000 additional WordNet-tagged verbs (Green et al., 2001a,b). The online *EVCA* index contains 3,024 verbs in 192 classes (with subclasses numbered from 9.1 to 57) – constituting a total of 4,186 verb entries. The LVD is significantly larger, with 4,432 verbs in 492 classes – a total of 10,003 entries.[7]

As in *EVCA*, each LVD class has a human-assigned name that generally refers to a prototypical verb of that class, e.g., Verbs of Contact: Touch Verbs, Run Verbs, and Verbs of Putting. The actual online form of the English entry *touch* is shown in Figure 5. This entry includes the root form of the word, its semantic class, a set of WordNet senses, thematic roles (clustered into what we will henceforth call a "grid"), and the LCS representation. Other entries in this class have a similar semantic structure – the same composition of LCS primitives – but vary in their semantic content or manner constant (indicated by `+ingly`).

```
(DEFINE-WORD
   :DEF_WORD "touch"
   :CLASS "20.a Verbs of Contact: Touch Verbs"
   :THETA_ROLES "_ag_th,loc,instr"
   :WN_SENSE (00820743 01832678 01455581)
   :LANGUAGE ENGLISH
   :LCS (act_on loc (* thing 1) (* thing 2)
        ((* [on] 23) loc (*head*) (thing 24))
        ((* with 19) instr (*head*) (thing 20)) (touch+ingly 26))
```

Figure 5. Lexicon Entry for *touch* in LVD Class 20.a *Verbs of Contact: Touch.*

```
(DEFINE-WORD
   :DEF_WORD "摩"
   :GLOSS "touch"
   :CLASS "20.a Verbs of Contact: Touch Verbs"
   :THETA_ROLES "_ag_th,loc,instr"
   :WN_SENSE (00820743 01832678 01455581)
   :LANGUAGE ENGLISH
   :LCS (act_on loc (* thing 1) (* thing 2)
        ((* [on] 23) loc (*head*) (thing 24))
        ((* with 19) instr (*head*) (thing 20)) (touch+ingly 26))
```

Figure 6. Lexicon Entry for *touch* in Chinese.

Our automatically acquired semantic lexicon includes LCSs for Chinese that are structurally analogous to their English counterparts. Figure 6 presents the result of building an online entry for the Chinese verb 摩 *mo*, the equivalent of *touch* in class 20.a.[8]

The LCS shown in each of these entries recursively associates logical heads with their arguments and modifiers. The logical head is represented as a primitive/field combination, e.g., $\text{ACT\_ON}_{\text{Loc}}$, or "`act_on loc`" in the actual online form. The arguments associated with this primitive/field combination are `(thing`

`1)` and `(thing 2)`. The numbers correspond to specific thematic roles in the `:THETA_ROLES` slot. For example, 1 corresponds to agent (i.e., `ag` in the online entry) and 2 corresponds to theme (i.e., `th` in the online entry). Both of these are obligatory arguments; thus, they are preceded by an underscore (_) in the `:THETA_ROLES` slot. In addition to these arguments, optional participants are indicated by commas (,). In this example, the optional participants are `loc` and `instr` – corresponding to numbers 23 and 19, respectively, in the LCS.

Thematic roles – coupled with the primitives of the LCS representation – facilitate the selection and ordering of target-language words in the ChinMT and MADLIBS systems described above. Once the LCS (i.e., the interlingua) is composed, the thematic roles – or, more accurately, their corresponding numbered positions – provide the basis for identifying the appropriate lexical entry and word order in the target language. The generator recursively checks whether all nodes in the composed LCS (including numbered positions) match those of candidate entries in the target language. The numbered positions are then further used to identify surface positions in the generated sentence, according to a "thematic hierarchy".[9]

The full set of roles used for our investigation is shown in Table I. An independent study indicates that many of these are widely agreed upon by human subjects (Habash, 2002). However, we expect that analogous (alternative) schemes – where roles are defined similarly for semantically related verbs – would be equally appropriate for our approach. An example of such a scheme is the one adopted in FrameNet (Baker et al., 1998), where the same role structure is associated with semantically similar verbs such as *argue* and *banter*. Additional thematic-role frameworks are discussed in Section 6.

An important contribution of the LVD – including its use of thematic roles at this level of granularity – is that it provides a systematic encoding of predictable role-to-LCS positions. The impact of this predictability is significant: it allows us to produce LCSs automatically for languages other than English – as long as we are able to identify a mapping from the roles specified for those languages to the LVD roles. The identification of such a mapping for Chinese and English is the core of the discussion in the next section.

3. Construction of Semantic Lexicon for Chinese–English MT

Mapping English thematic roles (e.g., LVD "Th(heme)") to their Chinese counterparts (e.g., HowNet "Patient") is the primary vehicle for creating links between Chinese and English verbs. This section demonstrates that it is possible to produce a lexicon by associating 709 Chinese HowNet concepts with 492 LVD classes, with a clear concept-to-class correspondence in a large majority of the cases. The class/role specifications in the resulting lexicon are used for automatic construction of LCS representations – by techniques presented in Dorr (1997b). The LCSs

Table I. Inventory of thematic roles in LVD.

Role	**Definition**	**Examples**
AG	Agent causing the event	John broke the vase. The hammer broke the vase.
TH	Affected or moved entity	John went to school. John broke the vase.
EXP	Experiencer of psych/mental event	John heard the vase shatter.
PERC	Perceived entity	He saw the play. He looked into the room. The cat's fur feels good to John.
INFO	Information conveyed	John memorized his lines.
SRC	Starting point of the event	John left the house. John ran away from home.
GOAL	Endpoint of the event	John turned into a monkey. John ran home. John ran to the store. John gave a book to Mary. John gave Mary a book.
BEN	Beneficiary of the event	John baked the cake for Mary. John baked Mary a cake. An accident happened to him.
PRED	Predicate or property	We considered him a fool. We pronounced him dead. She acted happy.
LOC	Location of event or thing	He lived in France. The water fills the box. This cabin sleeps five people. She grabbed him by the arm. She held the child in her arm. She coughed on John. The box on the shelf is red. She sang on the stage. The book unfolded before her.
PROP	Event or state	John wanted to go home. She imagined the movie to be loud.
POSS	Possessional predicate	John has five bucks. This box carries five eggs. This cabin sleeps five people.
PURP	Purpose or reason for event	He studied for the exam. He searched for rabbits.
MANNER	Manner of the main action	She went to school quickly.
INSTR	Instrument (modifier)	She hit him with a baseball bat.
MOD-PRED	Predicate or property (modifier)	The nation elected him president. They worshiped him as their leader. She imagined him as a prince.
MOD-POSS	Possessed item (modifier)	She bought it for five dollars. He loaded the cart with hay. He robbed him of his rights.

comprise the primary units of meaning in the interlingua for Chinese–English MT, as described above.

First, we introduce a Chinese conceptual database called HowNet. Next, we illustrate the magnitude of our resource-alignment techniques by comparing the content and size of the LVD and HowNet databases. Finally, we present our 3-step algorithm for mapping from English LVD thematic roles to their Chinese counterparts in the HowNet database.

3.1. HOWNET CONCEPTUAL DATABASE

HowNet is an on-line conceptual common-sense knowledge base that contains hierarchical information relating 80,000 Chinese words to 16,788 concepts.[10] Each HowNet entry was constructed by hand using a process that was entirely semantically motivated. For example, the HowNet word 佚 *yi* corresponding to the verb *lose* is associated with the HowNet concept |lose|失去. English translations – or "glosses" – are provided for each HowNet concept; however, English translations are not provided for any of the individual Chinese words.[11]

Our focus is on the verb hierarchy, which consists of 815 concepts, covering 16,647 Chinese verbs – a total of 20,467 verb-to-concept entries. The structure of the verb hierarchy is shown in Figure 7. The numbers assigned to each class are our own – these were labeled to make it easier for us to view the hierarchical structure of the database. Each number indicates the level of each concept in the verb hierarchy. We have excluded 106 HowNet verb concepts that are not associated directly with any Chinese words; these are "higher level" conceptual nodes with no direct Chinese realization (e.g., V.1 |static|). Thus, our investigation involves a total of 709 HowNet verb concepts.

Note that the highest two concepts in the verb hierarchy are "static" (V.1) and "act" (V.2). Under V.1, we find verbs such as 为 *wei* – the Chinese equivalent of *be*. Under V.2, we find verbs such as 开始 *kaishi* 'start'. The levels go much deeper than these, with the lowest ones at eight levels deep, e.g., V.1.2.1.6.3.3.1.15 for 痒 *yang* 'itch'.

Although synonym sets (called "synsets" in WordNet) are not explicitly defined in HowNet, they are implicitly encoded in the hierarchy. For example, the verbs 佚 *yi*, 遗失 *yishi*, and 遗 *yi* – all variants of the verb 'lose' – are associated with the same HowNet concept in the hierarchy: |lose|失去.

In addition to conceptual relations, HowNet developers have incorporated hand-constructed semantic-role specifications for each concept.[12] Consider the verb *cure*: this verb is associated with the semantic roles `(agent, patient, content, tool)`. In (9), the roles in the specification have the bindings as shown.

(9) The doctor cured the man of pneumonia with antibiotics.
`agent patient content tool`

Table II provides a list of the ten semantic roles used in HowNet.

V.1 |static|
V.1.1 |relation|
V.1.1.1 |isa|
V.1.1.2 |possession|
V.1.1.3 |comparison|
V.1.1.4 |suit|
V.1.1.5 |inclusive|
V.1.1.6 |connective|
V.1.1.7 |CauseResult|
V.1.1.8 |TimeOrSpace|
V.1.1.9 |arithmetic|
V.1.2 |state|
V.1.2.1 |StatePhysical|
V.1.2.2 |StateMental|

V.2 |act|
V.2.1 |ActGeneral|
V.2.1.1 |start|
V.2.1.2 |do|
V.2.1.3 |DoNot|
V.2.1.4 |Cease|
V.2.1.5 |Wait|
V.2.2 |ActSpecific|
V.2.2.1 |AlterGeneral|
V.2.2.2 |AlterSpecific|
V.2.3 |AlterRelation|
V.2.3.1 |AlterIsa|
V.2.3.2 |AlterPossession|
V.2.3.3 |AlterComparison|
V.2.3.4 |AlterFitness|
V.2.3.5 |AlterInclusion|
V.2.3.6 |AlterConnection|
V.2.3.7 |AlterCauseResult|
V.2.3.8 |AlterLocation|
V.2.3.9 |AlterTimePosition|

V.2.4 |AlterState|
V.2.4.1 |AlterPhysical|
V.2.4.2 |AlterStateNormal|
V.2.4.3 |AlterStateGood|
V.2.4.4 |AlterQuantity|
V.2.4.5 |AlterStateBad|
V.2.4.6 |AlterMental|
V.2.5 |AlterAttribute|:
V.2.5.1 |MakeHigher|
V.2.5.2 |MakeLower|
V.2.5.3 |AlterAppearance|
V.2.5.4 |AlterMeasurement|
V.2.5.5 |AlterProperty|
V.2.6 |MakeAct|:
V.2.6.1 |CauseToDo|
V.2.6.2 |CauseNotToDo|
V.2.6.3 |use|

Figure 7. HowNet verb hierarchy.

Our LVD–HowNet alignment, discussed next, involves the use of these semantic-role specifications for prioritization of candidate links between LVD and HowNet.

3.2. ENGLISH–CHINESE MAPPING: ALIGNMENT OF LVD WITH HOWNET

Our goal is to build LVD entries for 16,647 Chinese verbs from HowNet, starting with 10,003 LVD verb-to-class entries for 4,432 English verbs. Table III outlines the magnitude of the resources contributing toward this effort, including the number of verbs, classes (for *EVCA* and LVD), entries, and concepts (for HowNet).

We use a large (600k entry) Chinese–English dictionary called Optilex to link Chinese entries in HowNet to English entries in LVD.[13] Because the Chinese words in Optilex were not part-of-speech tagged in advance, we used automatic techniques to determine the part of speech for each entry (Olsen et al., 1998). The total number of verbs identified in Optilex (23,454) is 140% higher than the number of verbs included in HowNet (16,647). Analogously, the number of Chinese–English Optilex entries is higher than the number of verb-to-concept entries in HowNet: 43,840 compared to 20,467.

Our approach to linking LVD and HowNet entries relies primarily on single-word English glosses. There are 9,185 Optilex verbs with at least one single-word gloss (i.e., a total of 18,032 single-word-gloss entries). The remaining verbs contain only multi-word glosses, which are not directly linkable to LVD verbs, e.g., to be scattered and lost for the Chinese word 散失 *sanshi* in HowNet class |lose| 失去.

Table II. Ten most frequent among 69 HowNet roles.

Role	**Definition**	**Examples**
agent	Entity that acts to produce a result in the event	Birds fly. He bought a watch yesterday. His treatment of the data.
patient	Entity which is treated and typically changed in events of type *act*	They smuggled a lot of drugs. The computer was repaired by him. His treatment of the data . . .
content	Entity which is dealt with in the events, differing from *patient* in that it is not changed	I have been engaged in NLP for over 20 years. It started raining. She loves her children. I'm tired of your stupid conversation. Telephone numbers are difficult to remember. He said he would come. English is taught in that school.
experiencer	Main body in the events of type *state*	China's economy is developing rapidly these years. The plan to improve the quality of products failed. He was ill. He likes swimming. The man was disappointed. The competition will start soon.
target	Entity which is dealt with but not changed	Give me the book. We all respect him greatly. Who taught you chemistry. He spoke to her.
direction	Direction which an entity faces	The house faces the south Suddenly a car came up to me
LocationFin	Location where an entity locates after an event of type *AlterLocation*	They left for Tokyo. Please put them in the box.
LocationIni	Location where an entity locates before an event of type *AlterLocation*	He fled from the house. She will leave UK next week.
cause	Cause of an event	Why didn't he come yesterday. He died of illness. I am sorry he didn't come. Thanks for your prompt reply.
LocationThru	Location where an entity goes through during an event of type *AlterLocation*	He came to Moscow via London

Table III. Magnitude of resources used for LVD–HowNet alignment.

Resource	**# Verbs**	**# Sem Classes**	**# Verb Entries**	**# Concepts**
EVCA	3,024	192	4,186	–
LVD	4,432	492	10,003	–
HowNet	16,647	–	20,467	709
Optilex (single&multi-word glosses)	23,454	–	43,840	–
Optilex (single-word glosses only)	9,185	–	18,032	–
HowNet/Optilex	12,802	–	29,308	–

Table IV. Key components of LVD and HowNet.

English LVD	**Chinese HowNet**
Semantic classes with LCSs: Ex: 10.6.a Possessional Deprivation (cause X (go ident (toward ident Y (at ident Y cure+ed))))	**Concepts:** Ex: |Cure|
Words in class: Ex: 10.6.a absolve, cure, disabuse, exonerate	**Words with concept:** Ex: |Cure| treat, cure, heal
Thematic grid of roles with class: Ex: 10.6.a `(agent, theme, mod-poss)`	**Semantic roles with concept:** Ex: |Cure| `(agent, patient, content, tool)`

However, we are able to compensate for these multi-word glosses through gap compensation techniques (to be described in Section 4).

For our experiments, we start with the 12,802 Chinese verbs shared by HowNet and Optilex – a total of 23,260 Chinese–English entries. We supplement these Chinese verbs with 3,845 additional verbs from HowNet that are not found in Optilex. Thus, we consider a total of 16,647 verbs in this study.

The hierarchical structure of LVD – as an extension of *EVCA* – contrasts with the semantically motivated hierarchy of HowNet in that its design is based, in part, on syntactic behavior. Thus, we expect the alignment of these two differently motivated resources to provide insight into the validity of the hypothesis that "syntactic behaviors are fully semantically determined" (Levin, 1993).

On the other hand, we also expect the degree of success in linking the two resources to be higher than it would be in a strict *EVCA*-style approach because – as discussed in Ayan and Dorr (2002) – the LVD is more than a mere recasting of Levin's syntactically motivated classes. For example, the LVD is semantically motivated in its representation of aspectual distinctions (i.e., telicity). Such distinctions are determined by tests that do not rely on the syntactic behavior of verbs. An example is the use of a Dowty-style test for (a)telicity (Dowty, 1979): If the statement "*He was X-ing* entails *He has X-ed*" does *not* hold, then the verb *X* is telic (as in *run*); otherwise verb the *X* is atelic (as in *win*). Thus, the experiments described next serve to determine the degree of concept/class similarity of LVD and HowNet, given our enhancements to what otherwise would have been a purely syntactic-motivated framework.

Table IV highlights the key components of the HowNet and LVD resources. Our technique for mapping between English LVD classes and Chinese HowNet Concepts involves associating HowNet's semantic roles with LVD-based thematic roles. For example, the HowNet concept |Cure| is associated with the semantic roles `(agent, patient, content, tool)`, as in (9) above. The corresponding thematic grid in our LVD database is `(ag,th,mod-poss)`. These roles are associated with the first three noun phrases in the sentence; the fourth noun phrase would be considered a modifier and, thus, is not in the LVD grid. Although the HowNet and LVD roles are not in a one-to-one correspondence, they may be used

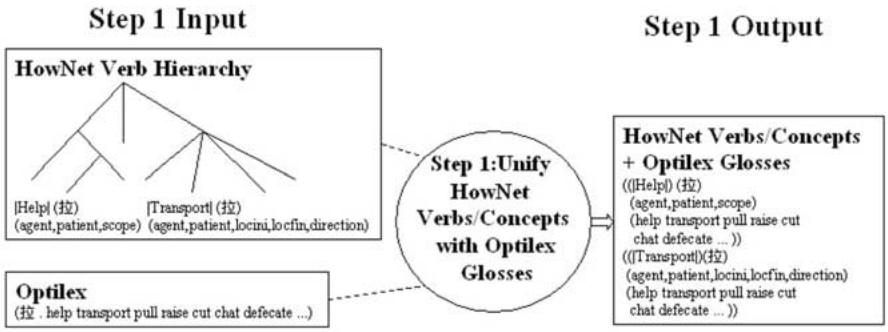

Figure 8. Unification of HowNet verbs/concepts with Optilex glosses.

for a "closest match" prioritization of candidate HowNet–LVD associations. This process consists of three steps, the details of which are described next.

3.3. STEP 1: UNIFY HOWNET VERBS/CONCEPTS WITH OPTILEX GLOSSES

The first step in aligning HowNet and LVD is the unification of 16,647 HowNet verbs and concepts with English glosses in Optilex. The procedure is illustrated in Figure 8. For each Chinese verb in the list of HowNet verbs, we extract candidate English glosses from Optilex. Since HowNet shares only 12,802 Chinese verbs with Optilex, the remaining 3,845 HowNet verbs are associated with a "dummy English gloss," *nil* (which is later overridden). The number of Chinese–English pairs resulting from this unification is 53,481. Each of these Chinese–English pairs is then associated with one or more of the 709 HowNet concepts (and their associated semantic roles) inducing 74,047 entries over all.

Consider the example given earlier of the multiply ambiguous Chinese verb 拉 *la*. This verb has several different Optilex glosses: *help*, *transport*, *involve*, *implicate*, *pull*, *drag*, *chat*, *defecate*, *slash*, *cut*, *raise*, *move*, *pressgang*. In HowNet, the verb is associated with multiple concepts: |Help|, |Transport|, |Include|, |Pull|, |Talk| |Excrete|, |Force|, |Attract|, and |Recreation|.

The combination of Optilex glosses with HowNet concepts is shown in Figure 12 for |Help| and |Transport|. The resulting entries are stored in a condensed 4-tuple form, `<Concept, Verb, HNRoles, Glosses>`, as shown in the block labeled "HowNet Verbs/Concepts + Optilex Glosses". Note that the semantic-role specification is carried along with each HowNet concept, i.e., `(agent, patient, scope)` and `(agent, patient, locini, locfin, direction)`, respectively. This information is used in later steps to create links between Chinese HowNet concepts and English LVD classes.

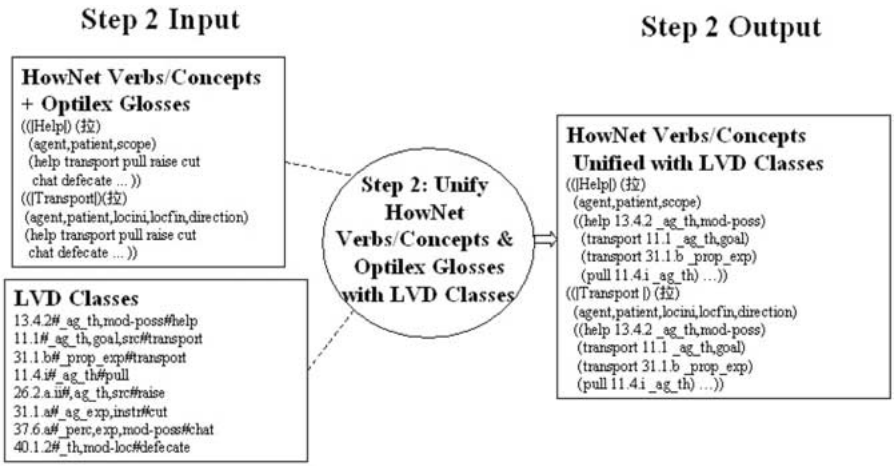

Figure 9. Unification of verbs/concepts/glosses with LVD classes.

3.4. STEP 2: UNIFY VERBS/CONCEPTS/GLOSSES WITH LVD CLASSES

The next step is to unify each entry of the verb/concept repository (the result of step 1) with the 492 LVD classes. The result is a larger repository of 88,049 entries, i.e., an average of 124 verb-to-class candidates per HowNet concept. This "fanout" of classes per concept is illustrated in our ongoing example, for the Chinese verb 拉 *la* in Figure 9. Entries produced by this step are an expanded version of the 4-tuples produced by the previous step, i.e., `<Concept, Verb, HNRoles, LVD-Classified-Glosses>`, as shown in the block labeled "HowNet Verbs/Concepts Unified with LVD Classes". The `LVD-Classified-Glosses` component of the 4-tuple is itself a triple of the form `<Verb, LVD-Class, LVD-Grid>`. This component allows us to distinguish different senses of the same verb. For example, the verb *transport* has two senses: one for the locational transfer sense (LVD class 11.1), and one for the emotional impact sense (LVD class 31.1.b).

The verb 拉 *la* is associated with 22 LVD classes, only some of which are shown in Figure 9: one for *help*, two for *transport*, one for *pull*, etc. Alphabetically, the full set of LVD classes identified for this verb is shown in Table V. These "candidate classes" are paired down to a smaller set of LVD classes by the next step.

3.5. STEP 3: ALIGN HOWNET WITH LVD VIA CLASS/ROLE-BASED RANKING

The final step involves the alignment of HowNet with LVD using class and role-based rankings. First, the Chinese–English entries associated with each HowNet concept are partitioned into groups whose English glosses correspond to translationally equivalent groups of verbs corresponding to LVD classes. This requires a ranking of candidate LVD classes for each HowNet concept. Those LVD classes that contain the largest number of English verbs matching the Optilex glosses are ranked highest.

Table V. 22 LVD classes associated with the Chinese verb *la*.

LVD Name	LVD Number	Examples
Admire:	31.2.b	*implicate*, *involve*
Amuse:	31.1.b	*cut*, *move*, *transport*
Braid:	41.2.2	*cut*
Breathe:	40.1.2	*defecate*
Build:	26.1.a	*cut*
Carry:	11.4.i	*carry*, *drag*, *pull*
Chitchat:	37.6.a	*chat*
Crane:	40.3.2	*raise*
Cut:	21.1.a	*cut*, *slash*
Cut:	21.1.d	*cut*
Equip:	13.4.2	*help*
Force:	12.a.ii	*pull*
Get:	13.5.1.a	*pull*
Grow:	26.2.a.ii	*raise*
Hurt:	40.8.3	*cut*, *pull*
Meander:	47.7.a	*cut*
Play:	009	*pawn*
Put:	9.4.a	*raise*
Search:	35.2.a	*drag*
Send:	11.1	*convey*, *ship*, *smuggle*, *transport*
Slide:	11.2.b	*move*
Split:	23.2.b	*cut*, *pull*

For example, one of the HowNet concepts associated with the Chinese verb 拉 *la* is |Transport|. This concept is associated with 60 additional Chinese verbs – the 61 verbs are associated with 201 translations occurring in 79 LVD classes. The LVD class associated with the highest number of translations (16 out of 201) is 11.1 Send, which includes *smuggle*, *transport*, *ship*, *convey*, etc. Thus, this class is given the highest ranking out of the 79 classes for the concept |Transport|.

Next, each candidate LVD class is prioritized according to the degree to which the corresponding thematic-role specification matches that of HowNet. For example, the second highest-ranking LVD class for the |Transport| concept is 31.1.b – as shown in the first lefthand block of Figure 15. This class corresponds to the Amuse Verbs, which is hypothesized due to the frequent occurrence of verbs expressing "emotional transfer" (i.e., verbs like *transport*, *move*, *cut*, as in (10)).

(10) The film moved me to tears.

However, this class is ruled out as a possibility for the |Transport| concept because the HowNet semantic roles `(agent, patient, locini, locfin, direction)` do not match the LVD thematic grid `_prop_ex` associated with the LVD class 31.1.b.

Table VI. Seed table for mapping HowNet roles into LVD roles.

Hownet	LVD Roles															
Roles	**ag**	**th**	**exp**	**goal**	**src**	**perc**	**info**	**pred**	**prop**	**loc**	**Instr**	**purp**	**mod-poss**	**mod-pred**	**ben**
agent	**278**	**77**	**32**	**1**	**2**	**3**	0	0	0	0	**4**	**11**	**7**	0	**4**
relevant	**15**	**4**	**4**	0	0	**1**	0	0	0	**1**	0	0	0	0	0
patient	0	**122**	**7**	**7**	0	**8**	0	0	0	0	0	0	0	0	0
possession	0	**28**	0	0	**1**	**2**	0	0	0	0	0	0	**3**	0	0
ContentProduct	0	**9**	0	0	0	0	0	0	0	0	0	0	0	0	0
experiencer	**13**	**32**	**33**	0	0	0	0	0	0	0	0	0	0	0	0
target	0	**7**	**12**	**27**	**1**	**17**	0	0	**3**	0	0	0	**2**	0	**1**
LocationFin	0	0	0	**31**	0	0	0	**1**	0	**8**	0	0	**2**	**2**	**1**
StateFin	0	0	0	**5**	0	**1**	0	0	0	0	0	0	0	0	0
source	0	**4**	0	0	**16**	0	0	0	0	0	0	0	0	0	**1**
LocationIni	0	0	0	0	**24**	0	0	0	0	**2**	0	0	0	0	0
StateIni	0	0	0	0	**4**	0	0	0	0	0	0	0	0	0	0
content	0	**31**	**1**	**2**	**2**	**14**	**20**	**3**	**6**	0	**3**	**3**	0	**1**	**1**
location	0	**1**	0	**1**	0	0	0	0	**1**	**6**	**2**	0	0	0	0
cause	0	0	0	**1**	0	**4**	0	**1**	**6**	0	**4**	**11**	**7**	**1**	0
partner	0	**2**	0	0	**3**	**3**	0	0	0	0	0	0	**11**	0	0
scope	0	**1**	0	0	0	**2**	0	0	0	**1**	**1**	0	**2**	**3**	0
contrast	0	**2**	0	**1**	0	**1**	0	0	0	0	0	0	**1**	0	0
purpose	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
range	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
beneficiary	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**4**
result	0	**1**	**1**	0	0	0	0	0	0	0	0	0	0	0	0
isa	0	**1**	0	0	0	**1**	**1**	0	0	0	0	0	0	0	0
partof	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0
manner	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

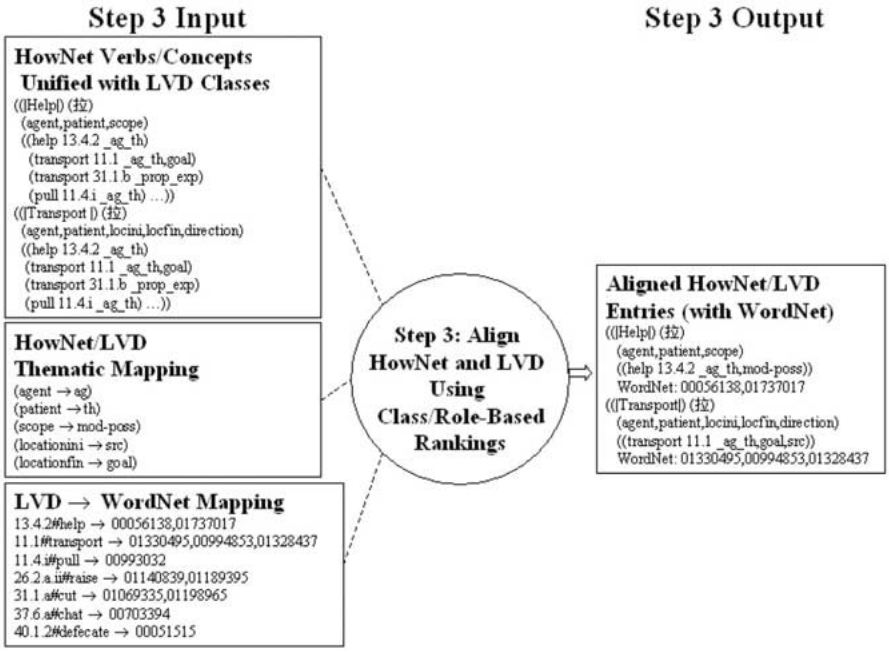

Figure 10. Alignment of HowNet with LVD via class/role-based rankings.

The role-matching procedure relies on Thematic Mappings (see the second lefthand block in Figure 10) which were automatically induced from a set of "seed mappings". A subset of the seed mappings are given in Table VI. These seed mappings were hand-constructed by a human at a rate of approximately 50 mappings per hour (7 hours) based on 350 randomly selected HowNet concepts. These were verified by a native Chinese speaker in a half day.

The numbers in the seed table indicate how many HowNet concepts correspond to a particular association between the HowNet roles and LVD roles specified in Tables I and II, respectively. These numbers are used for determining the weight of a match between a HowNet role and an LVD role. For example, 278 concepts correspond to an association between the HowNet role agent and the LVD role ag – the most heavily weighted association in the table. In general, the highest ranking HowNet–LVD role associations fall, roughly, along the diagonal of the seed table.

The matching procedure is simple: An LVD thematic grid is assigned a score corresponding to the sum of the numbers of all matching associations in the seed table. For example, the HowNet concept |Transport| is associated with the thematic-role specification `(agent, patient, LocationIni, LocationFin)` as in (11)

(11) John transported the goods from Boston to New York (westward).

This specification most closely matches the thematic grid `_ag_th,goal,src` – from LVD Class 11.1 Send – when 拉 *la* is translated as *transport*. This is because

the role associations `(agent->ag)`, `(patient->th)`, `(LocationIni->src)`, and `(LocationFin->goal)` correspond to the highest score: $278 + 122 + 24 + 31 = 455$. The second-highest score is significantly lower (32) – i.e., the LVD class 31.1.b Amuse, where the only non-zero association for the grid `_prop_ex` is `(agent->exp)`. Thus, the HowNet concept |Transport| is associated with the Send LVD class.

In the case of the HowNet concept |Help|, where 拉 *la* is translated as *help*, the semantic-role specification `(agent, patient, scope)` most closely matches the thematic grid `_ag_th,mod-poss` in the LVD Class 13.4.2 Equip – as in (12).

(12) John helped him with his work.

This is because the role associations `(agent->ag)`, `(patient->th)`, and `(scope->mod-poss)` correspond to the highest score: $278 + 122 + 5 = 402$. Thus, the HowNet concept |Help| is associated with the Equip LVD Class.

Once the procedure is complete, the English glosses associated with 拉 *la* are filtered down to the following:

- *help*: LVD Class 13.4.2 Equip; HowNet Concept |Help|
- *transport*: LVD Class 11.1 Send; HowNet Concept |Transport|
- *chat*: LVD Class 37.6.a Chitchat; HowNet Concept |Talk|
- *defecate*: LVD Class 40.1.2 Breathe; HowNet Concept |Excrete|
- *raise*: LVD Class 26.2.a.ii Grow; HowNet Concept |Include|
- *pull*: LVD Class 13.5.1.a Get; HowNet Concept |Recreation|

The Chinese verbs are additionally associated (for free) with senses from our previously-assigned WordNet 1.6 (Miller and Fellbaum, 1991; Fellbaum, 1998) – see the third lefthand block of Figure 10 – thus producing an Asian companion to the current (Euro)WordNet initiative. (More details about WordNet tagging are given in Dorr et al. (2000).) Because the WordNet tag has been added, the entries produced by this step are 5-tuples of the form `<Concept, Verb, HNRoles, LVD-Classified-Glosses, WN-Sense>`, as shown in the block labeled "Aligned HowNet/LVD Entries (with WordNet)" in Figure 10.

In our example, the WordNet senses that are assigned are senses 1 and 3 in the case of *help* (indexed as 01737017 and 00056138 in Figure 10) and senses 1, 2, and 4 in the case of *transport* (indexed as 01330495, 00994853, 01328437 in Figure 15):

- **help:**
  Sense 1: assist
  Sense 3: aid
- **transport:**
  Sense 1: transport
  Sense 2: carry
  Sense 4: send, ship

Note that the process described above relies on massive filtering of spurious class assignments. For example, the |Establish| HowNet concept is initially associated with 29 potential LVD classes, but it is ultimately associated with only two LVD classes, 29.2.c (Characterize) and 26.4.a (Create) due to our Class/Role ranking scheme. One example of an LVD class that was ruled out is the Change of State class, 45.4.a, associated with the Optilex translation *colonize* for the Chinese verb 殖民 *zhimin*. Although this is a perfectly valid LVD class assignment for the HowNet concept |Colonize|, it is not appropriate for the |Establish| HowNet concept. Because this class is ranked 8th for |Establish| – as opposed to 1st and 2nd place ranking for 29.2.c and 26.4.a, respectively – this assignment is ruled out by our algorithm.

Table VII characterizes the coverage of LVD with respect to the 709 HowNet concepts. For example, there are 111 HowNet concepts that are uniquely matched up to 1 LVD class, based on the class/role-based alignment scheme described above. Given that the majority of HowNet concepts are covered by 1 to 4 LVD classes – with only a small number of concepts represented by as many as 22 – we consider our algorithm to be a positive step in the direction of accurate HowNet–LVD alignment. However, our algorithm was only able to map 6,277 of the 16,647 HowNet verbs into LVD classes – a total of 8,089 entries – which is 43% of the total number of entries (18,530) that would be induced by "gap-compensation techniques", described next.

4. Compensating for Resource Deficiencies

The techniques described in the previous section create a bridge between entries in the Chinese HowNet conceptual hierarchy and the LVD semantic classes. However, the algorithm above is limited in that it relies heavily on the existence of single-word Chinese–English glosses in Optilex – and it also requires that these entries overlap with Chinese verbs in HowNet and English verbs in LVD. As it turns out, there are an additional 10,441 potential entries that can be extracted – for a total of 18,530 entries – if we apply techniques to compensate for the following gaps:

- **LVD Gaps**: the lack of LVD verbs corresponding to Optilex translations of Chinese verbs in HowNet (4,093 out of 16,647).
- **Optilex Gaps**: misspellings, omitted Chinese verbs, and lack of single-word glosses for Chinese–English entries (6,277 out of 16,647).

This section presents "gap compensation" techniques that allow us to produce a more complete alignment between HowNet and LVD. In order to induce this enhanced version of the algorithm, we built a new LVD-based resource called "Canonical Entries", which provides a canonical specification for each of the 709 HowNet concepts. First, we describe this new resource. Next, we describe how the resource is used to compensate for the two types of gaps above. Finally, we present the more complete results of our algorithm, using the canonical entries.

Table VII. Distribution of HowNet concepts across LVD classes.

# LVD	# HowNet
0	4
1	111
2	147
3	123
4	104
5	64
6	49
7	31
8	22
9	18
10	11
11	10
12	5
13	4
14	2
15	0
16	1
17	1
18	0
19	1
20	0
21	0
22	1

4.1. LVD-BASED CANONICAL ENTRIES

The entries in our new "Canonical Entries" database consist of pairs of LVD classes coupled with their associated prototype verbs. These canonical specifications provide a mapping between a HowNet concept and an LVD class/prototype-verb pair. That is, each HowNet concept is associated with a pair in the following form: `<Canonical-LVD-Class, Prototype-Verb>`.

In most cases, the prototype verb names the HowNet concept, e.g., *transport* for the |Transport| HowNet concept. In other cases – where the HowNet concept is not an English word – the prototype word is a realization of that concept, e.g., *belittle* for the |PlayDown| HowNet concept. A sample of the canonical specifications is given in Table VIII.

Candidate canonical specifications were automatically generated according to the highest ranking LVD class based on the "Thematic Mapping" approach described in Section 3.5 above. That is, we use the scores associated with the rankings for the 8,089 LVD-assigned Chinese entries to induce the most probable LVD

Table VIII. Sample of canonical specifications for filling resource gaps.

HowNet Concept	Canonical Specification
|Transport|	<11.1 Send, *transport*>
|BeNot|	< 22.2.a Amalgamate, *oppose*>
|Help|	< 13.4.2 Equip, *help*>
|Moisten|	< 45.4.a Change of State, *facilitate*>
|Excrete|	< 40.1.2 Breathe, *bleed*>
|Apologize|	< 32.2.a Long, *apologize*>
|PlayDown|	< 33.b Judgment, *belittle*>
|Naming|	< 29.3 Dub, *name*>
|Choose|	< 29.2.c, *choose*>
|Announce|	< 37.7.b Say, *announce*>
|Mean|	< 37.7.a Say, *signify*>
|Communicate|	< 37.9.c Advise *inform*>

class and canonical verb for a particular HowNet concept. Of course, for this, we require that each HowNet concept be associated with *at least one* English gloss that overlaps both Optilex and LVD. Although our results are even better when we have *more than one* overlapping English gloss, we do not require *every* Chinese HowNet verb to have an English translation. If we know *some* of the English translations for verbs associated with a given HowNet concept, we can make a reasonable guess about the translations of the remaining Chinese verbs. Thus, minimally, we need a bilingual dictionary that provides coverage of an element from each of the 709 verb classes (as Optilex does), but we do not require massive bilingual dictionaries for this effort.[14]

For example, the highest ranking LVD class for the |Transport| HowNet concept is 11.1 Send, which includes *smuggle*, *transport*, *ship*, *convey*, etc. This class has the highest number of verbs matching the English glosses associated with |Transport|; of these, the most common translation is the word *transport*. Thus, this LVD class is linked to the concept, coupled with the canonical verb *transport*, as shown in Figure 18. The full set of canonical assignments were hand-verified at a rate of 80 per hour for 709 classes.

We use these canonical specifications to compensate for gaps that arise in LVD and Optilex. For example, the Chinese verb 殡 *bin* has no Optilex gloss, yet it is associated with the |Transport| concept in HowNet. Thus, the canonical entry is automatically linked to this verb. The existence of this canonical entry ensures an accurate translation (*transport*) as well as an appropriate LVD-class assignment of 11.1 Send. This class assignment, in turn, corresponds to the thematic grid `_ag_th,goal,src`.

We now describe the use of the canonical entries for each of the resource gaps, in more detail.

4.2. LVD GAPS

An LVD gap is detected when an Optilex verb gloss for a Chinese verb does not occur in LVD. As mentioned above, 4,093 HowNet verbs have no LVD gloss. When this occurs, the canonical specification for the concept associated with the Chinese verb is automatically used to assign the verb an appropriate LVD class. For example, one Optilex gloss associated with the HowNet concept |Establish| (for the verb 重建 *chongjian*) is *reconstruct*, which does not occur in LVD. Our technique associates this Chinese verb with the canonical specification "29.2.c Characterize, *establish*," and the Chinese verb is then linked to the appropriate translation *establish*, providing a very specific meaning for *reconstruct* as in (13).

(13) The building was reconstructed as a memorial to those lost in the war.

An interesting byproduct of the handling of LVD gaps is that it allows us to enhance our LVD resource (and, additionally, the original *EVCA* index). For example the verb *reconstruct* can now be added to LVD Class 29.2.c, on a par with the previously classified LVD verb *establish*.

4.3. OPTILEX GAPS

An Optilex gap occurs when the translation of a Chinese verb is omitted – or its translation is a multi-word gloss or is misspelled. For example, the verb 摆布 *baibu* has only one Optilex gloss: *manipulate*. However, the word 摆布 is associated with two HowNet concepts, |Decorate| and |Control|. This gloss is only appropriate for the |Control| concept. The *decorate* meaning of 摆布 is omitted in Optilex.

Such gaps are detected by means of two types of information: (a) HowNet roles and LVD thematic grid; and (b) correlations between the gloss under question and *other* HowNet concepts. In this particular example, the thematic grid for *manipulate* in LVD is `(ag,exp,instr)`, which is ranked low (11th out of 28) with respect to the roles `(agent, patient)` associated with the HowNet concept |Decorate|. By contrast, this same LVD class has a high ranking (2nd out of 22) with respect to the HowNet |Control| concept due to a close match between `(ag,exp,instr)` and the HowNet thematic roles `(agent, patient, ResultEvent)`. In addition, the correlation of the gloss *manipulate* is much higher for HowNet's |Control| concept than it is for HowNet's |Decorate| concept (4 occurrences compared to 0). From these two types of information, we can conclude that the *decorate* sense of 摆布 *baibu* is missing from Optilex. As in the case with LVD gaps, our technique associates the Chinese verb with the canonical specification "9.8.b Fill, *decorate*" to compensate for this Optilex gap.

In addition to their usefulness in handling of gaps in our lexical resources, the canonical specifications proved useful for assigning LVD classes to Chinese verbs whose Optilex gloss is a multi-word gloss. For example, the Chinese verb 挨打 *aida* has only a single Optilex translation: *take a beating*. This verb is associated

Table IX. Distribution of HowNet concepts across LVD classes.

# LVD	# HowNet
0	2
1	84
2	143
3	132
4	116
5	64
6	56
7	32
8	24
9	18
10	9
11	14
12	3
13	6
14	1
15	1
16	1
17	0
18	1
19	1
20	0
21	0
22	1

with the HowNet concept |Suffer|, which has as its canonical specification "31.3.d Marvel, *suffer*". Thus, our technique associates the verb 挨打 with this canonical specification.

A similar approach is used for unknown or misspelled words. For example, the translation of 输送 *shusong* as in Optilex is misspelled as *tranport*. Because this verb is associated with HowNet's |Transport| concept, we associated this verb with the canonical specification "11.1 Send, *transport*".

4.4. APPLICATION OF GAP COMPENSATION

Using the gap compensation techniques described above, we have achieved a more refined HowNet–LVD mapping, providing an increase in LVD-classified Chinese words from the previous 8,089 entries to the current expanded set of 17,284 LVD-classified Chinese words.

Table IX characterizes the number of LVD classes required for coverage of 709 HowNet concepts. We considered this experiment to be a success for several reas-

ons: (a) In 359 cases (50% of the HowNet concepts), the partitioning corresponds to three or fewer LVD classes; (b) Most HowNet concepts with two or more partitions had a very heavy association with a single LVD class (60% or higher), with most other partitions falling around 20% or lower; (c) Only two cases did not correspond to any LVD class (i.e., degenerate HowNet concepts for which no correlations with LVD could be found); (d) There were virtually no partitionings (a handful of single HowNet concepts) exceeding 13 LVD classes.

5. Results

Above we described our HowNet–LVD mapping – supplemented with gap compensation techniques – which yielded 17,284 LVD-classified Chinese words. We now turn to two quantitative evaluations of our approach. First, we will examine the effect of different translation resources on the mapping of HowNet classes to LVD classes. Second, we will compare the results of our LVD-based assignment to a set of manually assigned HowNet verbs.

5.1. EFFECT OF TRANSLATION RESOURCES ON HOWNET–LVD ALIGNMENT

At the time of our initial experiment, the HowNet resource did not include English translations. Although the translation resource we used was the Optilex dictionary, our technique was developed to accommodate an arbitrary translation resource for mapping between HowNet concepts and LVD classes. The most recent release of HowNet associates English glosses with each word in a class. To assess the impact of this additional translation resource, we performed a three-way comparison, performing the HowNet-LVD mapping with: (a) Optilex translations alone; (b) HowNet translations alone; and (c) a merged resource including translations from both HowNet and Optilex.

These mappings differed only in the Chinese–English gloss pairs available to the mapping process. However, since our mapping process relies on English single-word glosses to link HowNet classes to LVD classes, several differences in translation resources can impact the mapping results: (a) no translations for Chinese word in the resource, (b) no single-word translation for Chinese word in the resource, and (c) different single-word translation(s) for a Chinese word in the resource. Thus, a complete quantitative evaluation should include the same three-way comparison, both with and without gap compensation techniques.

We compute precision and recall measures for HowNet-LVD mapping – for each of the individual resources – both relative to the merged resource and relative to each other. Specifically, the *precision* of Resource A relative to Resource B is computed as the number of correct LVD-assigned pairs in Resource A relative to Resource B divided by the total number of assignments in Resource A (14). The *recall* of Resource A relative to Resource B is computed as the number of

Table X. Precision and recall for HowNet–LVD mappings (with and without canonical grid information).

Contrast	Precision		Recall	
	w/o Canon	w/ Canon	w/o Canon	w/ Canon
HowNet vs Optilex	0.61 (1264/2073)	0.65 (1537/2351)	0.46 (1264/2729)	0.55 (1537/2793)
Optilex vs HowNet	0.42 (1264/3018)	0.51 (1537/3032)	0.61 (1264/2070)	0.65 (1537/2351)
HowNet vs Merged	0.79 (1636/2073)	0.82 (1918/2351)	0.61 (1636/2674)	0.67 (1918/2844)
Optilex vs Merged	0.71 (2137/3018)	0.75 (2273/3032)	0.79 (2137/2712)	0.80 (2273/2844)

correct LVD-assigned pairs in Resource A relative to Resource B divided by the total number of assignments in Resource B (15).

(14) $$\text{Precision(A vs B)} = \frac{\#\text{ Correct Assignments in Resource A wrt Resource B}}{\#\text{ Total Assignments in Resource A}}$$

(15) $$\text{Recall(A vs B)} = \frac{\#\text{ Correct Assignments in Resource A wrt Resource B}}{\#\text{ Total Assignments in Resource B}}$$

We also assess the impact of our gap compensation techniques, i.e., using the Canonical Entries resource to aid in the assignments. The effect of the Canonical Entries is measured – in the precision and recall formulas above – by adding each canonical entry to the set of assigned mappings obtained by each resource, when it was not automatically generated. The results appear in Table X. In this table, the basic assignment – without the use of the Canonical Entries – is referred to as "**w/o Canon**"; the enhanced assignment – using the Canonical Entries – is referred to as "**w/Canon**".

We find that the HowNet resource achieves higher precision, as might be expected since the available translations are limited to those the designer believed appropriate for each class. The Optilex resource achieves higher recall by drawing from a wider variety of alternate translations. Note that our gap compensation techniques provide an improvement of all measures – as much as 21% higher in some cases (e.g., Optilex vs. HowNet) – and smooths differences between the resources.

If we further examine the translations used to make these assignments, we find 7,653 Chinese-word–English-gloss pairs in common, 17,609 pairs from HowNet, and 14,252 from Optilex, from a total of 24,205 assigning pairs. The results indicate that a merged translation resource, drawing from both HowNet and LVD/Optilex, can produce a richer and more robust mapping among the concept classes. For example, the HowNet concept |WeatherChange| is associated with three verbs, 下雨 *xiayu* 'rain', 下雪 *xiaxue* 'snow', and 普降 *pujiang* 'fall all over the area'. Whereas the first two verbs have translation equivalents that link directly into our thematic-grids (and, hence, our WordNet senses), the third verb is WordNet

linked solely by virtue of our thematic-grid matching routine. This routine allows us to determine that the closest English equivalent for 普降 *pujiang* is *precipitate* – a verb that does not show up in the HowNet hierarchy. Thus, our integration of LVD/Optilex with the HowNet resource has provided a more comprehensive linking to thematic grids and WordNet senses than would be available in either resource alone.

5.2. COMPARISON TO MANUAL LVD CLASS AND THEMATIC-GRID ASSIGNMENT

In addition to the comparisons of HowNet–LVD alignment for different translation resources described above, we performed a quantitative analysis of our automatic algorithm relative to manual lexicon creation. We compare the results of our LVD-based assignment to a set of manually assigned HowNet verbs. Two Chinese language experts provided us with LVD-based assignments for a set of 272 separately hand-tagged Chinese verbs. These verbs provided complete verbal coverage for a set of ten articles in the economic domain from the Xinhua News Agency that formed the development set for the ChinMT project (Dorr et al., 1998; Olsen et al., 1998). Manual assignment resulted in 1,188 distinct thematic-grid labels and 1,282 LVD class labels.

For these verbs, our automatic algorithm proposes 577 class/grid assignments, some of which are duplicates. We report precision and recall measures for both class and grid assignments relative to manual assignment. To reiterate, precision is the ratio of the number of correct automatic assignments to the total number of automatic assignments, whereas recall is the ratio of the number of correct automatic assignments to the total number of assignments according to the "gold standard" manual labeling. Note that duplicate assignments of LVD class and grid to each Chinese word are counted only once in the computation of recall. Correctness is defined according to the relaxed agreement criteria specified below.

We also contrast the corresponding measures computed for two plausible naive strategies as baselines: (a) random assignment of an LVD class and grid label to each Chinese–English entry associated with a HowNet concept; and (b) assignment of the most frequently occurring LVD class and grid label to each Chinese–English entry associated with a given HowNet concept. An example of this second strategy is the one given earlier for the HowNet concept $|Decorate|$: the LVD class "9.8 *Fill*)" is the most frequently occurring class associated with the English glosses for this concept, so this class is naively assigned to the entire HowNet concept. For these contrastive runs, we replace each of the 577 automatic LVD-based assignments with a corresponding LVD-based assignment generated either randomly or by the most frequent label.

Our criteria for agreement between manual and automatic assignments are as follows:

Table XI. Precision and recall of our HowNet–LVD alignment.

Criterion	Precision	Recall
LVD Grid - Auto	0.62 (359/577)	0.25 (269/1083)
LVD Grid - Random	0.10 (58/577)	0.00 (53/1083)
LVD Grid - Most Freq	0.36 (206/577)	0.07 (79/1083)
LVD Class - Auto	0.63 (363/577)	0.29 (279/946)
LVD Class - Random	0.04 (20/577)	0.02 (19/946)
LVD Class - Most Freq	0.18 (104/577)	0.04 (41/946)

– LVD classes are said to agree if a supercategory is assigned by the human, e.g., 40.7.ii.a and 40.7.i both match the human-assigned LVD Class 40.7.
– Thematic grids agree if the same roles appear in the same order, without regard for obligatory versus optional distinctions, e.g., `_ag,th` matches `_ag_th`.

These results appear in Table XI, where "Auto" refers to our automatic algorithm, "Random" refers to the first naive baseline, and "Most Freq" refers to the second naive baseline.

We achieve precision of approximately 0.62 for both LVD class and thematic-grid assignment; recall levels are lower, at approximately 0.24. As illustrated in the table, the automatic technique we have developed substantially outperforms either random or most frequent LVD class assignment and random thematic-grid assignment. While still large, the contrast with most frequent thematic-grid assignment is less dramatic. The relatively good performance of the most frequent thematic-grid assignment is easily explained by the fact that 28% of the verbs can appear as the basic transitive, the most common thematic grid. Thus, the transitive grid assignment produces precision of 0.36.

The relatively lower numbers for recall are best understood in terms of two features. First, this technique is more focused on high precision than on recall. Second, the "majority rules" strategy for selection among alternative likely assignments will tend to prefer more common class assignments, when many, less frequently acceptable class assignments are available.

6. Related Work of Others

Although the translation of English semantic lexicons into other languages has proven difficult (Jones et al., 1994; Nomura et al., 1994; Saint-Dizier, 1996), regularities can be found in some online resources (Dang et al., 1998; Dorr and Jones, 1999; Olsen et al, 1998). We exploit such regularities in a framework that is similar in nature to the "intersective-class" approach of Dang et al. (1998),[15] with the following extensions:

– The construction of an entry for all *EVCA* verbs – plus those in the enhanced LVD – rather than a small set of verbs (the *break* class)

– The provision of a thematic-role based filter for a more refined version of verb-class assignments
– Concept alignment across two different language hierarchies (Chinese and English) rather than one
– Mappings between Chinese and English thematic roles
– Hooks into WordNet 1.6 senses for both languages.[16]

The work of Carpuat et al. (2002) suggests that corpus-based methods are superior to those that utilize existing semantic resources because they do not rely on structural information in the ontology. Although we rely on the structure of the English classification system of Dorr (2001), we take this structure to be a constant across all language pairs, i.e., it is not redeveloped for each language pair. Moreover, we do not require bilingual corpora, unlike alternative approaches, in keeping with the generation-heavy philosophy adopted for many of our multilingual applications. However, we do require that the foreign-language semantic resource include language-independent concepts – a standard requirement of all lexical-acquisition approaches. We also require that the resource include thematic roles or argument frames, which are now becoming increasingly available in many new online resources including the latest releases of WordNet (Fellbaum, 1998) and PropBank (Palmer et al., 2001).

Our choice of thematic roles coincides with those agreed upon by a wide range of human subjects (Habash, 2002), although this choice is not without controversy. Thematic role definitions range from very specific versions – domain-specific roles such as FROM_AIRPORT (Stallard, 2000 and Hobbs et al., 1997) or linguistically-specific roles such as "X-er" (where X is the name of the verb) (Kingsbury and Palmer, 2002) – to more general "proto-roles" or "macroroles," such as PROTO-AGENT (Dowty, 1991 and vanValin 1993). Our roles lie somewhere in between. That is, each role unifies all occurrences of a particular position associated with a large collection of similarly structured LCSs – but we do not generalize these further (e.g., we do not have a notion of "patient", which might serve as a collapsed version of theme, possessed, and perceived). A cogent review of alternative representations for thematic roles (also called "semantic roles") is provided by Gildea and Jurafsky (2002).

7. Conclusion and Future Work

We have presented an approach to aligning two large-scale online resources, HowNet and LVD. The lexicon resulting from this approach is large-scale, containing 18,530 Chinese entries. The technique for producing these links involves matching thematic grids in HowNet with those in LVD. Our results indicate that the correspondence is very high between the 709 Chinese HowNet concepts and the 492 LVD classes. In our comparisons with manual lexicon creation, our automatic techniques were shown to achieve 62% precision, compared to a much lower precision of 10% for arbitrary assignment of semantic links. Thus, we see our

techniques as the first step toward a general approach to building repositories for interlingual-based NLP applications.

Our work has shown that it is possible to combine different types of knowledge from existing resources in ways that improve upon the coverage and robustness of each of these independent resources. One area of investigation that has allowed us to enrich the existing resources is the development and application of gap compensation techniques, allowing us to fill in possible Chinese–English links where none existed previously.

Until last year, HowNet contained no English translations. Thus, our initial experiments used Optilex to produce candidate English translations. In the latest version of HowNet (Dong, 2000), the English translations are included; however, our work has provided the basis for increasing recall – acquisition of thousands of correct Chinese–English entries that do not currently exist in HowNet – and, moreover, it has provided a link into the semantic classes underlying a large English conceptual database. Since the new HowNet was released, we have been able to execute a more accurate evaluation of our Chinese–English links – in particular, we use the English translations in HowNet to determine the precision of our approach (overall accuracy of the Chinese–English links we have already automatically acquired). Finally, given that our initial work did not make use of the English translations in HowNet, we expect those same techniques to be generally applicable to *other* foreign language semantic hierarchies where English translations are not available. We predict this will occur more and more frequently, as online (non-bilingual) linguistic resources continue to be made available in multiple languages (see, for example, Hovy (1998)).

One area of future work is the use of our gap compensation techniques to enhance foreign-language resources such as the HowNet hierarchy. For example, in some cases, the HowNet hierarchy incorrectly associates a Chinese word with a particular concept. This is the case for the two Chinese verbs 扎花 *zhahua* and 绣花 *xiuhua*, which are associated with the |Decorate| concept. These two verbs are both translated as *embroider* and would be more appropriately associated with the |Weave| concept. We may be able to detect such discrepancies by means of LVD-class frequency for a particular HowNet concept. In the current example, only 2 out of the 17 verbs associated with HowNet's |Decorate| concept (the two miscategorized Chinese verbs above) are associated with an LVD class that is not 9.8 (Fill) or 9.9 (Butter). Ultimately, the miscategorized verbs should be disassociated from the HowNet concept upon detection of this discrepancy.

We are currently using the lexicon for word-sense disambiguation in MT and cross-language information retrieval. As we saw above the Chinese verb 拉 *la* has several possible translations, but not all of these will be appropriate in every context. If we can determine which HowNet concept corresponds to 拉 *la*, then we will translate it appropriately. For example, if the HowNet concept is |Transport|, the translation would be *ship* or *transport*, but not *slash*, *chat*, *implicate*, etc. We can detect which HowNet concept is appropriate by examining the other words

in the sentence. If those words co-occur with *other* Chinese verbs associated with a particular HowNet concept (as determined through a corpus analysis), then it is likely that that HowNet concept is the appropriate one for the Chinese verb. That is, if we find other verbs from a given HowNet concept occurring in the same context, then we can hypothesize that this particular verb has the meaning of this HowNet concept.

The algorithm for mapping between HowNet concepts and LVD classes requires human intervention – i.e., the seed mappings given earlier. However, it is possible to automate the construction of a ranked mapping between thematic grids by counting correspondences between LVD-based roles and the HowNet-based roles across the entire concept space. This approach is also currently under investigation.

Another area of investigation is the use of a WordNet-based distance metric (e.g., the information-content approach of Resnik (1995)) for additional pruning power in the HowNet–LVD alignment. Because each of the entries in the LVD classification is associated with a WordNet sense, it is possible to rule out certain class assignments for a given HowNet concept by examining semantic distance between the Optilex glosses for a particular Chinese word and the glosses for other words associated with that concept.

Acknowledgements

The University of Maryland authors are supported, in part, by PFF/PECASE Award IRI-9629108, DOD Contract MDA904-96-C-1250, DARPA/ITO Contract N66001-97-C-8540, and NSF CISE Research Infrastructure Award EIA0130422. Dekang Lin is supported by Natural Sciences and Engineering Research Council of Canada grant OGP121338. We are indebted to Nizar Habash, Maria Katsova, and Scott Thomas for their assistance with experimental runs on the data and their useful commentary and aid in the preparation of this document, and to James Allan for help with the Inquery configuration.

Notes

[1] Presentation by Charles Wayne, Darpa TIDES PI Meeting, San Diego, CA, July, 2001.

[2] The ambiguity in the word 拉 *la* can often be resolved if it is combined with other characters. For example, 拉车 *la che* unambiguously means 'pull a cart'. However, since object dropping is a frequent phenomenon in Chinese, it is not uncommon for verbs like *la* to appear without an argument that easily disambiguates the word. To accommodate this, we allow for multiple possibilities in the lexicon.

[3] The system has since been upgraded to incorporate the Halogen generator (Langkilde-Geary, 2002).

[4] Note that this representation of the surface sentence does not include the ON and WITH components shown below in Figure 5 since there are no modifier phrases such as *on the tail* or *with a comb* in this particular example.

[5] The *EVCA* index may be found at ftp://linguistics.archive.umich.edu/linguistics/texts/indices/evca93.index.

[6] In the work of Green et al. (2001a,b), the LVD resource was referred to as "Levin+". It was subsequently renamed because its former name implied that the database constitutes a simple extension to Levin's original work, rather than a complete overhaul.

[7] The 492 classes are assigned more specific numbers than those in the original *EVCA* index. For example, the *EVCA* class "51.3.2" of *Run Verbs* is sub-divided into "51.3.2.a.i", "51.3.2.a.ii", "51.3.2.b.i", "51.3.2.b.ii", "51.3.2.c", and "51.3.2.d" according to certain aspectual distinctions (Dorr and Olsen, 1996; Olsen et al., 1997a,b).

[8] The English "gloss" is not used during the MT process; we store it in lexical entries for convenience only (i.e., readability of the lexicon by English speakers).

[9] Details of the thematic hierarchy are omitted here – see Dorr et al. (1998a) for more details. Briefly, the surface elements are generated in the target language according to the following hierarchical ordering: ext > ag > instr > th > perc > goal > src > ben.

[10] Available at http://www.keenage.com/html/e_index.html.

[11] The latest version of HowNet does associate English translations with Chinese words. We discuss this point further below.

[12] Independent verification of the HowNet roles by a native speaker indicates that, at least for a large cross-section of 100 entries, these roles are generally internally consistent (Tiejun Zhao, p.c., 2002).

[13] Optilex is a machine-readable version of the CETA dictionary licensed from the MRM Corporation, Kensington, MD.

[14] This makes the approach easily extensible to new language pairs – a bilingual lexicon containing *hundreds*, but not necessarily *thousands*, of entries should be sufficient.

[15] Intersective classes refer to the grouping together of subsets of existing Levin-based classes with overlapping members. For example, *cut*, *tear*, and *split* occur together in more than one of Levin's semantic classes (Change of State Verbs, as in *The bread cuts/tears/splits easily*, and Split Verbs, as in *She cut/tore/split the bread apart*), so they are grouped together in an intersective class.

[16] The WordNet hooks are currently undergoing a mapping from WordNet 1.6 to the up-to-date WordNet 1.7 (Ken Litkowski, p.c., 2002).

References

Ayan, N. F. and B. J. Dorr: 2002, 'Generating A Parsing Lexicon from an LCS-Based Lexicon', in *LREC 2002 Workshop Proceedings: Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Las Palmas, Spain.

Baker, C. F., C. J. Fillmore, and J. B. Lowe: 1998, 'The Berkeley FrameNet Project', in *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Quebec, pp. 86–90.

Ballesteros, L. and W. B. Croft: 1997, 'Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval', in *SIGIR '97: Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, pp. 84–91.

Carpuat, M., G. Ngai, P. Fung, and K. Church: 2002, 'Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet', in *Proceedings of the 1st Global WordNet Conference*, Mysore, India.

Dang, H. T., K. Kipper, M. Palmer, and J. Rosenzweig: 1998, 'Investigating Regular Sense Extensions Based on Intersective Levin', in *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Quebec, pp. 293–299.

Dong, Z.: 1988a, 'Enlightment and Challenge of Machine Translation', *Shanghai Journal of Translators for Science and Technology* **1**, 9–15.

Dong, Z.: 1988b, 'Knowledge Description: What, How and Who?', in *Proceedings of International Symposium on Electronic Dictionary*, Tokyo, Japan, p. 18.

Dong, Z. D.: 1988c, 'MT Research in China', in Dan Maxwell, Klaus Schubert and Toon Witkam (eds), *New Directions in Machine Translation*, Foris, Dordrecht, pp. 85–91.

Dong, Z.: 2000, 'HowNet Chinese–English Conceptual Database', Technical Report Online Software Database, Released at ACL. http://www.keenage.com.

Dorr, B. J.: 1993, *Machine Translation: A View from the Lexicon*, MIT Press, Cambridge, MA.

Dorr, B. J.: 1994, 'Machine Translation Divergences: A Formal Description and Proposed Solution', *Computational Linguistics* **20**, 597–633.

Dorr, B. J.: 1997a, 'Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring', in *Fifth Conference on Applied Natural Language Processing*, Washington, DC, pp. 139–146.

Dorr, B. J.: 1997b, 'Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation', *Machine Translation* **12**, 271–322.

Dorr, B. J.: 2001, 'LCS Verb Database', Technical Report Online Software Database, University of Maryland, College Park, MD. http://www.umiacs.umd.edu/∼bonnie/LCS_Database_Docmentation.html.

Dorr, B. J., N. Habash, and D. Traum: 1998, 'A Thematic Hierarchy for Efficient Generation from Lexical-Conceptal Structure', in Farwell et al. (1998), pp. 333–343.

Dorr, B. J. and D. Jones: 1999, 'Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision', in E. Viegas (ed.), *Breadth and Depth of Semantic Lexicons*, Kluwer Academic Publishers, Norwell MA, pp. 79–98.

Dorr, B. J. and M. Katsova: 1998, 'Lexical Selection for Cross-Language Applications: Combining LCS with WordNet', in Farwell et al. (1998), pp. 438–447.

Dorr, B. J., G.-A. Levow, D. Lin, and S. Thomas: 2000, 'Chinese–English Semantic Resource Construction', in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece, pp. 757–760.

Dorr, B. J., M. A. Martí, and I. Castellón: 1997, 'Spanish EuroWordNet and LCS-Based Interlingual MT', *Proceedings of the Workshop on Interlinguas in MT, MT Summit*, San Diego, CA, pp. 19–32.

Dorr, B. J. and M. B. Olsen: 1996, 'Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization', *Machine Translation* **11**, 37–74.

Dorr, B. J., L. Pearl, R. Hwa, and N. Habash: 2002, 'DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment', in Richardson (2002), pp. 31–43.

Dowty, D.: 1979, *Word Meaning in Montague Grammar*, Reidel, Dordrecht.

Dowty, D.: 1991, 'Thematic Proto-Roles and Argument Selection', *Language* **67**, 547–619.

Farwell, D., L. Gerber, and E. Hovy (eds): 1998, *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas, AMTA'98*, Springer, Berlin.

Fellbaum, C.: 1998, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Gildea, D. and D. Jurafsky: 2002, 'Automatic Labeling of Semantic Roles', *Computational Linguistics* **28**, 245–288.

Green, R., L. Pearl, B. J. Dorr, and P. Resnik: 2001a, 'Lexical Resource Integration across the Syntax-Semantics Interface', in *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*, Pittsburg, PA, pp. 71–76.

Green, R., L. Pearl, B. J. Dorr, and P. Resnik: 2001b, 'Mapping WordNet Senses to a Lexical Database of Verbs', in *Association for Computational Linguistics 39th Annual Meeting and 10th Conference of the European Chapter*, Toulouse, France, pp. 244–251.

Habash, N.: 2000, 'Oxygen: A Language Independent Linearization Engine', in John S. White (ed.), *Envisioning Machine Translation in the Information Future: 4th Conference of the Association for Machine Translation in the Americas, AMTA2000*, Springer, Berlin, pp. 68–79.

Habash, N.: 2002, 'IL Annotation Experiment', in *Workshop on Interlingual Reliability, Fifth Conference of the Association for Machine Translation in the Americas, AMTA2002*, Tiburon, CA.

Habash, N. Y.: 2003, 'Generation-Heavy Hybrid Machine Translation', Ph.D. thesis, Department of Computer Science, University of Maryland, College Park, MD.

Habash, N. and B. Dorr: 2001, 'Large Scale Language Independent Generation Using Thematic Hierarchies', in *MT Summit VIII: Machine Translation in the Information Age*, Santiago de Compostela, Spain, pp. 139–144.

Habash, N. and B. J. Dorr: 2002, 'Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation', in Richardson (2002), pp. 84–93.

Hobbs, J. R., D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson: 1997, 'FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text', in E. Roche and Y. Schabes (eds), *Finite-State Language Processing*, MIT Press, Cambridge, MA, pp. 383–406.

Hovy, E.: 1998, 'Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses', in *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, Spain.

Hull, D. A. and G. Grefenstette: 1996, 'Experiments in Multilingual Information Retrieval', in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR '96*, Zurich, Switzerland, pp. 49–57.

Jackendoff, R.: 1983, *Semantics and Cognition*, MIT Press, Cambridge, MA.

Jackendoff, R.: 1990, *Semantic Structures*, MIT Press, Cambridge, MA.

Jones, D., R. Berwick, F. Cho, Z. Khan, K. Kohl, N. Nomura, A. Radhakrishnan, U. Sauerland, and B. Ulicny: 1994, 'Verb Classes and Alternations in Bangla, German, English, and Korean', Technical report, Massachusetts Institute of Technology.

Kingsbury, P. and M. Palmer: 2002, 'From Treebank to PropBank', in *LREC 2002: Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, pp. 1989–1993.

Langkilde, I. and K. Knight: 1998a, 'Generating Word Lattices from Abstract Meaning Representation', Technical report, Information Science Institute, University of Southern California.

Langkilde, I. and K. Knight: 1998b, 'Generation that Exploits Corpus-Based Statistical Knowledge', in *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Quebec, pp. 704–710.

Langkilde, I. and K. Knight: 1998c, 'The Practical Value of n-Grams in Generation', in *Proceedings of the 9th International Natural Language Generation Workshop (INLG '98)*, Niagra-on-the-Lake, Ontario.

Langkilde-Geary, I.: 2002, 'An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator', in *International Natural Language Generation Conference (INLG '02)*, Marriman, NY.

Levin, B.: 1993, *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, IL.

Levow, G.-A., B. J. Dorr, and D. Lin: 2000, 'Construction of Chinese–English Semantic Hierarchy for Cross-Language Retrieval', in *Proceedings of the Workshop on English-Chinese Cross Language Information Retrieval, International Conference on Chinese Language Computing*, Chicago, IL, pp. 187–194.

Miller, G. A. and C. Fellbaum: 1991, 'Semantic Networks of English', in B. Levin and S. Pinter (eds), *Lexical and Conceptual Semantics*, Blackwell, Cambridge, MA, pp. 197–229.

Nomura, N., D. A. Jones, and R. C. Berwick: 1994, 'An Architecture for a Universal Lexicon: A Case Study on Shared Syntactic Information in Japanese, Hindi, Bengali, Greek, and English', in *COLING 94: The 15th International Conference on Computational Linguistics*, Kyoto, Japan, pp. 243–249.

Oard, D. W.: 1998, 'A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval', in Farwell et al. (1998), pp. 472–483.

Oard, D. W. and B. J. Dorr: 1996, 'A Survey of Multilingual Text Retrieval', Technical Report UMIACS TR 96-19, CS-TR-3615, University of Maryland, Institute for Advanced Computer Studies. http://www.glue.umd.edu/~oard/research.html.

Olsen, M. B., B. J. Dorr, and D. J. Clark: 1997a, 'Using WordNet to Posit Hierarchical Structure in Levin's Verb Classes', in *Proceedings of the Workshop on Interlinguas in MT, MT Summit*, San Diego, CA, pp. 99–110.

Olsen, M. B., B. J. Dorr, and S. C. Thomas: 1997b, 'Toward Compact Monotonically Compositional Interlingua Using Lexical Aspect', in *Proceedings of the Workshop on Interlinguas in MT, MT Summit*, San Diego, CA, pp. 33–44.

Olsen, M. B., B. J. Dorr, and S. C. Thomas: 1998, 'Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese', in Farwell et al. (1998), pp. 41–50.

Palmer, M., A. Joshi, M. Marcus, M. Liberman, and F. Pereira: 2002, 'Multilingual PennTools', TIDES Presentation, University of Pennsylvania.

Palmer, M. and J. Rosenzweig: 1996, 'Capturing Motion Verb Generalizations with Synchronous Adjoining Grammars', in *Expanding MT Horizons, Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, Quebec, pp. 76–85.

Palmer, M., J. Rosenzweig, and S. Cotton: 2001, 'Automatic Predicate Argument Analysis of the Penn TreeBank', in *Human Language Technologies Conference*, San Diego, CA.

Palmer, M., J. Rosenzweig, and H. T. Dang: 1997, 'Intersective Levin Classes', in *Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C. Presentation at the Working Group on Combining Knowledge Sources for Automatic Semantic Tagging.

Palmer, M. and Z. Wu: 1995, 'Verb Semantics for English-Chinese Translation', *Machine Translation* **10**, 59–92.

Peters, W., P. Vossen, P. Diez-Orzas, and G. Adriaens: 1998, 'Cross-Linguistic Alignment of Wordnets with an Inter-Lingual-Index', *Computers and the Humanities* **32**, 221–251.

Procter, P.: 1978, *Longman Dictionary of Contemporary English*, Longman, London.

Resnik, P.: 1995, 'Using Information Content to Evaluate Semantic Similarity in a Taxonomy', in *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95*, Montréal, Québec, pp. 448–453.

Richardson, S. D. (ed.): 2002, *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002*, Springer, Berlin.

Saint-Dizier, P.: 1996, 'Semantic Verb Classes Based on 'Alternations' and on WordNet-like Semantic Criteria: A Powerful Convergence', in *Proceedings of the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases*, Toulouse, France, pp. 62–70.

Stallard, D.: 2000, 'Talk'n'Travel: A Conversational System for Air Travel Planning', in *Association for Computational Linguistics 6th Applied Natural Language Processing Conference*, Seattle, Washington, pp. 68–75.

van Valin, J. R. D.: 1993, 'A Synopsis of Role and Reference Grammar', in J. Robert D. van Valin (ed.), *Advances in Role and Reference Grammar*, John Benjamins, Amsterdam, pp. 1–164.

Viegas, E., B. A. Onyshkevych, V. Raskin, and S. Nirenburg: 1996, 'From Submit to Submitted via Submission: On Lexical Rules in Large-Scale Lexicon Acquisition', in *34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, pp. 32–39.

Vossen, P.: 1998, *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht.

Vossen, P., L. Bloksma, A. Alonge, E. Marinai, C. Peters, I. Castellon, A. Marti, and G. Rigau: 1998, 'Compatibility in Interpretation of Relations in EuroWordNet', *Computers and the Humanities* **32**, 153–184.

Vossen, P., P. Diez-Orzas, and W. Peters: 1997, 'The Multilingual Design of EuroWordNet', in *Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Application*, Madrid, Spain.

Weinberg, A., J. Garman, J. Martin, and P. Merlo: 1995, 'Principle-Based Parser for Foreign Language Training in German and Arabic', in J. K. Melissa Holland and M. Sams (eds), *Intelligent Language Tutors: Theory Shaping Technology*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 23–44.