Translating English and Mandarin Verbs with Argument Structure (Mis)matches using LCS Representation

Mari Broman Olsen
UMIACS
University of Maryland
College Park, MD 20742
phone: +1 (301) 405-6754
fax: +1 (301) 314-9658
molsen@umiacs.umd.edu

Abstract

This paper applies and evaluates a semi-automatically acquired Mandarin Chinese lexicon (Olsen et al., 1998) with respect to translation of English and Chinese verbs in a UNESCO text (Otero, 1997). I demonstrate how Lexical Conceptual Structure templates allow the same semantic structure to apply both to verbs with thematic roles incorporated in the verb itself, and those requiring external thematic complements. Using as examples the English verb *provide*, the Chinese counterpart 提供, and the English counterparts of 提供 in the text, I show how potential translations are included or eliminated automatically based on their thematic role structure. The example illustrates (i) how an interlingual thematic representation based in large part on English argument structure may be adapted felicitously to a historically unrelated language, and (ii) how an interlingual (IL) resource developed for analysis may also be used in generation.

Keywords: lexical conceptual structure, lexicon, machine translation, theta roles, thematic structure, interlingual MT

1 Introduction

In (Olsen et al., 1998) we report on a refinement to our procedure for porting lexical conceptual structure (LCS) to new languages. We describe how we related verb-complement information in the definition field of an on-line Chinese-English dictionary to Lexical Conceptual Structures (LCS) based on English. The creation of the verbs lexicon is part of a larger project for analyzing Chinese and generating English using the LCS as an interlingua. Quantitative analysis of the effect of these changes are difficult at this point: the project is still in its early stages, and a 'control' system is not available for comparison. However, the qualitative analysis of verbs in the workshop text (Otero, 1997) (available at http://crl.nmsu.edu/Events/FWOI/SecondWorkshop/text.html) suggests that this is a promising direction for representing cross-linguistic variability in predicate argument structure in an interlingual framework, for generating Chinese as well. This paper suggests how the verb lexicon resource may be used for selecting and eliminating translation candidates, and how the performance might be enhanced.

Section 2 outlines the role of the thematic structure in our interlingua (IL). Section 3 describes the theoretical underpinnings and procedure for creating the Chinese verb lexicon. Section 4 applies this resource to generating and eliminating candidate translations of the English verb *provide*, the Chinese counterpart 提供, and the English counterparts of 提供 in the workshop text (Otero, 1997).

2 Thematic Structure: Grids

Thematic ("theta") role structure is the interface between the syntactic component (parsing) and the lexical-semantic component, the Lexical Conceptual Structure (LCS). Verbs that appear in the same types of sentences, with the same syntactic and semantic complement types are assigned to the same class. The syntactic and semantic behavior of the class is abbreviated in the form of thematic grids, a form of semantically tagged subcategorization frame.

In thematic grids, theta roles preceded by an underscore '_' are obligatory, and those preceded by a comma ',' are optional. Parentheses '()' indicate that the role must be expressed as a complement of a preposition or complementizer (e.g. the infinitival *to* in English). If a form is specified, that preposition or complementizer must be the head of the phrase. For example, the thematic grid `ag_th,src(from),goal(to)` indicates that agent and theme are obligatory, that source and goal are optional, and that they must be expressed (respectively) by *from* and *to* prepositional phrases, if they appear. Assigning `ag_th,src(from),goal(to)` to the *Send* verbs, for example (class 11.1 in (Levin, 1993)), allows these verbs to appear in sentences like (1)-(4), but not (5)-(8), since the obligatory theme argument is missing.

(1) I sent the book.

(2) I sent the book to Mary

(3) I sent the book from the warehouse.

(4) I sent the book from the warehouse to Mary.

(5) ∗ I sent.

(6) ∗ I sent to Mary.

(7) ∗ I sent from the warehouse.

(8) ∗ I sent from the warehouse to Mary.

The thematic roles map directly into numbers, representing variables in the LCS. Although theta roles are theoretically unordered (Rappaport and Levin, 1988), the numbers correspond to a "canonical" linear position in a sentence and relative structural height in syntax and LCS trees. Thus `1` in the LCS corresponds to the `ag`(ent) thematic role and `2` to `th`(eme), since agents usually precede themes and occur higher in the syntactic tree: in a sentence with both agent and theme, the agent will typically be the subject and the theme the object, with both preceding other arguments.

The LCS for the above grid (simplifying irrelevant details) is given below: agent = `thing 1`, theme = `thing 2`, source preposition = `thing 3`, source complement of the preposition = `thing 4`, goal preposition = `thing 5`, goal complement of the preposition = `thing 6`. The $*$ markers indicate where arguments are instantiated.

```
(cause (* thing 1)
  (go loc (* thing 2)
      ((* to 5) loc (thing 2) (at loc (thing 2) (thing 6)))
      ((* from 3) loc (thing 2) (at loc (thing 2) (thing 4))))
  (!!-ingly 26))
```

The grids therefore group verbs by "semantic structure" (Levin and Rappaport Hovav, 1995). In contrast to "semantic content"—the idiosyncratic aspect of verb meaning—semantic structure determines syntactic patterning within and across languages (Dorr and Oard, 1998; Dorr and Katsova, 1998; Grimshaw, 1993; Pinker, 1984; Pinker, 1989). Most importantly for our system, the thematic grid or set of grids assigned to a verb class maps directly into interlingual LCS structures (Dorr et al., 1995).

In our system, both parsing and semantic analysis key off the thematic structure of the verb, covering as much of the source language (SL) sentence as possible using the largest grid assigned to a verb as a member of a class. The analysis module would first try to match the structures in the full grid `ag_th,src(from),goal(to)`, matching a *from* or *to* prepositional phrase to source and goal, respectively, in sentences like (4). Only if the relevant phrases were not found, as in (1), would the algorithm assign `_ag_th` alone to a sentence.

The generation module selects a target language (TL) structure that matches as much of the SL LCS as possible. If the SL grid is available for the head verb in the TL, the analogous structure is generated. If the structures diverge, other heuristics apply, for example covering part of the LCS in the main clause, and the remainder in additional clauses (see, e.g. (Dorr and Olsen, 1996)).

3 Methodology

3.1 Theoretical background

In creating LCS entries for new languages, we leverage the fact that semantic structure overlaps across languages to a large degree. That is, the building blocks of the LCS are valid cross-linguistically. Furthermore, the semantic structure of particular verb classes are also constant within and across languages. For example, verbs meaning 'to fill' involve three entities in their semantic structure: someone or something doing the filling, a container that becomes filled/full and the contents of the container; more generally, an `ag`(ent), a `th`(eme) and a `mod-poss`, a "possessed" item.[1] Verbs meaning 'to carpet, to cloak,' and 'to plug' have similar semantic structures. In English, all the entities (represented in the grid in (9)) may be overt complements, as in (10). Alternatively, specified arguments may be left implicit, such as the `mod-poss` in (11)

[1]`Mod-poss` roles are paraphraseable as complements of *have*, with the theme as the subject: *The hole has dirt (in it)*.

(9) `_ag_th,mod-poss(with)`

(10) Derek filled the hole with dirt.

(11) Derek filled the hole.

Implicit entities are still present in the meaning of the sentence: *Derek* is filling the bucket with something in (11), although it is not specified. Thus, (10) and (11) are generated by the same LCS representation for *fill* in English.

Alternatively, the two sentence-types may be represented by two different verbs in a language: one which expresses the `mod-poss` role as a complement, whereas another incorporates ("lexicalizes") the role. Furthermore, the exact contents of the lexicalized role may be part of the meaning of the verb. Chinese verb entries from the Chinese-English Translation Assistance (CETA) dictionary exhibit a great degree of this type of variation.[2] Entries in (12)-(14) are glossed with simple English verbs (given with information from CETA's Pinyin, definition, and simplified character fields). Entries in (15)-(17) have more specific requirements on elements in their semantic structure. For example, although (12) and (15) have the same LCS, only (12) may have a theme complement; the emperor is the theme of (15) and is expressed as part of the verb.

(12) force 迫使 po4_shi3

(13) run 奔跑 ben1_pao3

(14) walk 走 zou3

(15) force_the_sovereign_to_abdicate 逼宫 bi1_gong1

(16) run_around_spreading_the_news 奔走相告 ben1_zou3_xiang1_gao4

(17) walk_dragging_one's_feet 擦拉着走 ca1_la5_zhe5_zou3

Cross-linguistic variation may be expressed analogously. For example, English and Chinese verbs may differ only in terms of which semantic structure complements are lexicalized in the verb and which expressed overtly. It is this insight which drives the automatic portion of our Chinese thematic grid construction.

3.2 Automatic grid generation

Each thematic grid in the candidate set for a given Chinese verb describes the argument structure for the head verb of the English gloss from CETA. For instance, the candidate set for the Chinese verb in (15) above, glossed 'to force the sovereign to abdicate,' contains the grid `_ag_th,prop(to)`, because the English verb *force* takes an agent, theme and optional propositional complement. After parsing the gloss into subphrases, we posit that 'the sovereign' is theme, and 'to abdicate' the propositional element, assuming that a gloss of this sort means the theme and propositional element are lexicalized in the Chinese verb and *not* expressed as overt complements. The grid is therefore reduced to `_ag`, 'the sovereign' and 'to abdicate' are inserted directly into the LCS for the Chinese entry, and the following grid is submitted for manual correction:

[2] CETA's 600k Chinese-English entries were compiled from some 250 dictionaries, some general purpose, others domain-specific or bilingual (Russian-Chinese, English Chinese, etc.). The CETA group, started in 1965 and continuing into the present decade, was a joint government-academic project. The machine-readable version of the CETA dictionary, *Optilex*, licensed from the MRM corporation, Kensington, MD.

(18) 002 `_ag` 逼宫 bi1_gong1 force_the_sovereign_to_abdicate (th = sovereign) (prop = to_abdicate)

Similarly, 填土 tian2_tu3 receives the grid shown in (10), but with the `mod-poss(with)` lexicalized by the verb, and thus removed from the grid:[3]

(19) 9.8 `_ag_th` 填土 tian2_tu3 fill_in_with_earth (mod-poss = earth)

The output of automatic generation and manual checking is a (set of) LCS representations for each subsense of a each verb selected from CETA.[4] In the next section I illustrate how the Chinese LCS database will be used in translation between English and Chinese. Creation of the verb entries is part of a larger project to create a usable lexicon for interlingual machine translation from Chinese into English. Since manual checking of the grids is still in process, I discuss only automatically generated grids. Additional steps remain in creating the final LCSs from the manually-verified grids, such as incorporating into the LCS material in the definition that does not saturate a thematic role.

4 Examples

This section outlines how our interlingual LCS format provides appropriate tools for generating translation equivalents between English and Chinese, even those differing in argument structure. Since the representation is interlingual, it can aid in translation independent of which language is source and which target. When Chinese is the source language, the Chinese lexicon aids in filling out the complements of the English verbs. When English is the source language, the Chinese lexicon aids in lexical selection. For purposes of evaluation, the texts are treated as potential translations of each other, although both are translations from a Spanish text (Otero, 1997).

Consider the English verb *provide*, the Chinese counterpart 提供, and the English counterparts of 提供 in the text, selected because of their frequency in the text as well as the variation in complement structure shown by their counterparts in the other language. In particular, the English verb *provide* is similar to the LIGHT VERBS, such as *do, give, have, make, take, let*, that derive their event semantics in large part from their complements. In many cases, constructions with light verbs and NPs and the verbal form of the NP are mutual paraphrases:

(20) give a groan = groan

(21) take a walk = walk

(22) have a drink = drink

(23) provide service/help = serve/help

Section 4.1 discusses the examples in which *provide* and 提供 are counterparts of each other in the text. Section 4.2 deals with mismatches: when 提供 is translated with other English verbs, as well as when *provide* could be translated with other Chinese verbs.

4.1 Headword matches

Provide occurs four times in the English version of the text.

[3] See (Olsen et al., 1998) for further discussion of how gloss and LCS elements are matched.

[4] Certain verbs and senses were eliminated by our language experts because they were classical Chinese, or from extremely specialized sources; see(Olsen et al., 1998).

(24) ACCION International is a U.S.-based private non-profit organization that currently *provides* technical assistance to a network of institutions in thirteen countries in Latin America and six cities in the United States.

(25) The latter brought with them leadership and seed capital, while the former *provided* technology and methodology.

(26) PRODEM, as the programme was named, *provided* credit and training to broaden employment opportunities for the very poor self-employed, encourage investment in microbusinesses, and increase the income generated by this sector.

(27) The enormous demand, coupled with PRODEM's desire to *provide* savings services to its borrowers and to access capital markets for funds, moved PRODEM's leadership towards the transformation of this non-profit institution into a fully chartered private commercial bank specializing in microfinance–the first in the world.

In all cases, *provide* corresponds to the following Chinese verb:

(28) HWS: 提供
PY : ti2 gong1
STC: 2251 0180
DEF: to provide, to furnish, to offer (e.g., ideas, materials, aid)

The automatic portion of the thematic grid acquisition program creates three separate subentries, generating candidate grids from the English head verbs *provide, furnish*, and *offer*.[5] Each set of grids corresponds to a subentry or "sense" of the verb. Grids are prefaced by the class number to which they are assigned, as derived from (Levin, 1993).

(29) *provide*
13.4.1.a `_ag_th,goal(to)`
13.4.1.b `_ag_th,mod-poss(with)`

(30) *furnish*
13.4.1.a `_ag_th,goal(to)`
13.4.1.b `_ag_th,mod-poss(with)`

(31) *offer*
13.3 `_ag_th,goal(to)`
29.2.c `_ag_th_pred(as)`
29.2.d `_ag_th_prop(to)`
48.1.2 `_ag_th,ben(to)`

In the examples above, *provide* has the following thematic structure.

(32) `ag` = a U.S.-based private non-profit organization
`th` = technical assistance
`goal` = to a network of institutions ...

(33) `ag` = The former
`th` = technology and methodology

[5] Actually, since the entry had no REF field in CETA, it was not identified as coming from the 20-or so modern dictionaries from which we gleaned our verb list. These grids are therefore hand-created in accordance with the algorithm. The verb has subsequently been added to the list.

(34) `ag` = PRODEM
`th` = credit and training
`purp` = to broaden employment opportunities ...

(35) `ag` = PRO
`th` = savings services
`goal` = to its borrowers

Since the `purp`(ose) role is a modifier, it can be attached outside every LCS structure. Thus, the above examples may all be expressed by the grid assigned to *provide* as a member of the Verbs of Fulfilling class: `_ag_th,goal(to)`, representing obligatory agent and theme, with an optional goal in a *to* prepositional phrase.[6]. Neither *furnish* nor *offer* occurs in this corpus; I will therefore focus on the role of the `_ag_th,goal(to)` grid, with *provide* as a headword in the remainder of the paper.

The Chinese and English LCS structures match in cases (25)-(27): both have the headword *provide* in their representation and are therefore assigned (at least) the same grids. For translating Chinese 提供 ti2_gong1 into English, *provide* will be among the possible translations, as will *furnish* and *offer*. Selection among them uses information outside the LCS, for example text frequency in particular domains.[7] As seen below, the two alternate translations are acceptable.

(36) ACCION ...currently *furnishes/offers* technical assistance to a network of institutions ...

(37) ... the former *furnished/offered* technology and methodology.

(38) ... PRODEM ... *furnished/offered* credit and training ...

(39) ... PRODEM's desire to *furnish/offer* savings services to its borrowers.

For translations from English *provide* to Chinese, 提供 ti2_gong1 will also be one of a set of possible translations that match the headword *provide* and the relevant grid. A large set of verbs is excluded, even though their glosses are headed by *provide*, because the theme role is saturated in the verb.[8]

(40) 13.4.1 . `_ag` 反◆ fan3_zheng4 provide_evidence_to_the_contrary (th = evidence) (goal = contrary)

(41) 13.4.1 . `_ag,goal(to)` 出资 chu1_zi1 provide_money (th = money)

(42) 13.4.1 . `_ag,goal(to)` ◆济 zhou1_ji4 provide_material_assistance (th = material_assistance)

(43) 13.4.1 . `_ag,goal(to)` 伴舞 ban4_wu3 provide_musical_accompaniment_for_dancing
(th = musical_accompaniment)

(44) 13.4.1 . `_ag,goal(to)` 包饭 bao1_fan4 provide_or_receive_meals_for_fixed_time_and_price (th = meals)

(45) 13.4.1 . `_ag,goal(to)` 包伙 bao1_huo3 provide_or_receive_meals_for_fixed_time_and_price (th = meals)

(46) 13.4.1 . `_ag,goal(to)` 管饭 guan3_fan4 provide_food (th = food)

(47) 13.4.1 . `_ag,goal(to)` 管住 guan3_zhu4 provide_lodging (th = lodging)

(48) 13.4.1 . `_ag,goal(to)` 救济 jiu4_ji4 provide_relief (th = relief)

[6] The roles in (33) and (34) may also be generated by the grid `_ag_th,mod-poss(with)`, with the optional `mod-poss` absent

[7] In our current system, selection among LCS synonyms is part of the generation module, based on Nitrogen, a prototype system from the Information Sciences Institute (University of Southern California) that combines symbolic phrase structure rules with a bigram model of English derived from corpora (Knight and Hatzivassiloglou, 1995)

[8] A diamond indicates no glyph mapping is available for the character.

(49) 13.4.1 . _ag,goal(to) 救灾 jiu4_zai1 provide_disaster_relief (th = disaster_relief)

(50) 13.4.1 . _ag,goal(to) 穷鸟入怀 qiong2_niao3_ru4_huai2 provide_relief_for_the_unfortunate (th = relief)

(51) 13.4.1 . _ag,goal(to) 身先士卒 shen1_xian1_shi4_zu2 provide_leadership (th = leadership)

(52) 13.4.1 . _ag,goal(to) 以工代赈 yi3_gong1_dai4_zhen4 provide_work_as_a_form_of_relief (th = work)

(53) 13.4.1 . _ag,goal(to) 荫庇 yin4_bi4 provide_shade (th = shade)

The following verbs are also excluded, since they lack an agent role.

(54) 47.8 . _th_loc 备 bei4 be_provided_or_equipped_with

(55) 47.8 . _th_loc 备有 bei4_you3 be_provided_with

Several candidate translations remain. The following are glossed simply 'provide' (possibly with other subsenses) and therefore have the requisite argument structure to match the English LCS.

(56) 13.4.1 . _ag_th,goal(to) 备办 bei4_ban4 provide

(57) 13.4.1 . _ag_th,goal(to) 发放 fa1_fang4 provide

(58) 13.4.1 . _ag_th,goal(to) 给 ji3 provide

(59) 13.4.1 . _ag_th,goal(to) 供 gong1 provide

(60) 13.4.1 . _ag_th,goal(to) 具 ju4 provide

(61) 13.4.1 . _ag_th,goal(to) 资 zi1 provide

In these cases, heuristics outside the LCS interlingua would again be necessary to select among candidates.[9] For example, more frequent strings in a relevant data set could be preferred, as well as frequency of particular senses, as can be derived from WordNet for English, for example. In addition, all else being equal, multiple-character strings may be preferred over the more polysemous single characters (John Kovarik, p.c.). Under the latter diagnostic 提供 would be preferred to (59), which has only the second of the two characters.

The remaining cases to be excluded as inappropriate translations involve entries with *for* followed by an NP. Some of these appear to contain material that could saturate a goal or theme role, but were not so analyzed, because the grids required a bare NP theme or a *to*-PP goal.[10] In (62)-(64), the *for* appears to head a goal-like prepositional phrase.

(62) 13.4.1 . _ag_th,goal(to) 生救 sheng1_jiu4 provide_for_oneself_by_production (manner = by_production)

(63) 13.4.1 . _ag_th,goal(to) 自备 zi4_bei4 supply_or_provide_for_oneself

(64) 13.4.1 . _ag_th,goal(to) 自理 zi4_li3 provide_for_oneself

In (65)-(66), the *for* appears to be a particle — part of a complex verb — with a theme complement.

[9] We do not currently have a Chinese generation module that makes use of such information.

[10] Complementizer *for* (followed by a VP) saturates a purp(ose) thematic role, if purp is present in the English grid (Olsen et al., 1998).

(65) 13.4.1 . `_ag_th,goal(to)` 赡 shan4 provide_for_the_daily_needs

(66) 13.4.1 . `_ag_th,goal(to)` 赡养 shan4_yang3 provide_for_the_daily_needs

Examples (67)-(69) contain idioms that have neither theme nor goal NPs.

(67) 13.4.1 . `_ag_th,goal(to)` 未雨绸缪 wei4_yu3_chou2_mou2 provide_for_a_rainy_day

(68) 13.4.1 . `_ag_th,goal(to)` 不时之需 bu4_shi2_zhi1_xu1 provide_for_a_rainy_day

(69) 13.4.1 . `_ag_th,goal(to)` 利用厚生 li4_yong4_hou4_sheng1
provide_for_the_well-being_of_the_people

Incorporating thematic material in the NPs following *for* would require more fine-grained analysis of the verbs in Levin (Levin, 1993), since not every verb in the class with *provide* permits *for* as a particle or a goal preposition:

(70) Goal preposition:
∗ He entrusted for his mother a piece of land.
?∗ I credited for myself a $200 deposit.
? I furnished for myself a room.
? I issued for myself a security.
(also *leave present serve supply trust*)

(71) Particle + Theme:
I left for him my entire estate.
I served for him my favorite meal.
? I supplied for him the best produce.
∗ I credited for him a $200 deposit.
(also *entrust furnish issue present serve trust*)

Some refinement can be done manually, incorporating the material into the LCS for the grids that survive manual checking. The resulting material could then be used in the refining the English verb classes.

Thus, our Chinese LCS database aids in generation for both Chinese and English, as demonstrated by the cases in the text for which the headword (constant) in the LCSs match. When the SL is English, the Chinese lexicon aids in lexical selection by filtering out possible entries with the same headword that do not match in thematic structure, since some roles are lexicalized in the verbs themselves. Selection among the alternatives that remain, requires heuristics outside the LCS. In addition, other candidate translations could be generated, given a systematic way of relating the contents of an overt complement with the contents of a saturated role, via something like WordNet. The current system accounts for selectional preference in an opaque way, as part of the corpus-based information bundled in the Nitrogen generator. When SL is Chinese, the Chinese lexicon aids in filling out the complements of the English verbs. Again, a way of measuring the relationship between overt complements and lexicalized roles is required.

4.2 Headword mismatches

In the six remaining cases where 提供 appears, the English counterpart would require enhancement to the existing system, some of which could be incorporated in the LCS.[11] Our English generation system would again select *provide, offer, furnish* as candidates, and again these are legitimate translations.

[11] There are no cases in the text where *provide* does not correspond to 提供.

Three times in the text the Chinese verb takes as object (or passive subject) an NP incorporating money: a 'loan' or 'loan portfolio'.

(72) Its network of eighteen independent organizations in Latin America has *lent* over $1 billion to microenterprises in the last five years, in loans averaging less than $500.

(73) In its first five years of operation, PRODEM *financed* loans to over 13,300 microentrepreneurs, 77 per cent of whom were women, disbursing over $27 million in loans averaging $273.

(74) BancoSol's outstanding portfolio is some $35 million, about one fourth of which is *funded* through savings deposits.

Our current system would generate English *provide* (or *furnish*, or *offer*) with a similar NP. However in the English text version, three other verbs are used, with strong selectional preference for themes lexicalizing the concept MONEY. *Finance* and *lend* are both members of the same class (13.1) and assigned the grid `_ag_th_goal(to)`, and *fund* is (13.4.2) `_ag_th,mod-poss(with)`. As mentioned above, both of these grids are also assigned to 提供 ti2_gong1.

(75) *lend* (= (72)):
提供了１０亿多美元的贷款
'provide/lend 1 billion more than US$ of loan'

(76) *finance* (= (73)):
提供了融资贷款
'provide has financed loan'

(77) *fund* (= (74)):
大约四分之一是 ... 提供的
'about fourth of [BancoSol's portfolio] is ... provided'

Although the English verbs in these examples appear to incorporate a MONEY theme (e.g. *finance/lend/fund* = 'provide money'), these verbs still allow an overt theme, as in the examples in (72), (73), and (74). Furthermore, the relation between the verb and the theme in these cases is selectional preference, rather than lexicalization: indeed, themes without the MONEY concept are permitted:[12]

(78) lend some help/a car; finance a house/a play; fund a van/the vacation

Selectional restriction does, however, play a role in lexical selection: verbs with strong selectional relationships of a particular type could be substituted for by verbs with incorporated themes of the same type. Indeed, 出资 chu1_zi1 in (41) above is also an appropriate Chinese translation of the verbs in (72), (73), and (74) (John Kovarik, p.c.). In our current system, these type of relationships are only (and perhaps accidentally) accounted for using the statistical information bundled in the modified Nitrogen generator.

Incorporating this information into the LCS would require identifying and relating appropriate selectional features, a non-trivial task (Resnik, 1996). For example, if 出资 chu1_zi1 were glossed according to its more specific meaning 'provide capital', and *finance* had the selectional restriction MONEY, the lexical selection algorithm would need to know that CAPITAL and MONEY were related,

[12] Perhaps money *is* a lexicalized theme for *finance* and *fund*, with the object argument is a purpose. Thus, *finance,* , for example, would have the thematic grid `_ag_purp` (theme = money (or finances or funds)).

via WordNet, for example, in which CAPITAL ISA ASSET, and MONEY ISA MEDIUM OF EXCHANGE, which ISA ASSET. A measure of the sort in (Resnik, 1996) would be appropriate.

Measuring these relationships relates to the more general problem of selectional restrictions. I have considered only verbs that appear to completely saturate the thematic roles within the verb. However, it is often possible, particularly in English, to either override the content lexicalized in the verb, as in (79), or to further specify it, as in (80). In these cases, the `mod-poss` appears to be lexicalized in the verb, but a *with*-PP complement is permitted with the same role, provided it adds additional information.

(79) Gordy buttered his bread with margarine.

(80) Sylvia papered her dining room with antique wallpaper.

Languages may differ on this parameter, in terms of whether an additional thematic role may be specified. The degree of difference/similarity required between the role in the verb and that expressed in a complement may also be appropriately measured via something like WordNet.

For the remaining three examples, it would be useful in two of these cases to recognize a correspondence between the object of the light verb and a main verb usage:

(81) Other shareholders included ACCION, Calmeadow from Canada, which had *been* very *instrumental in* the formation of the bank, Fundes from Switzerland and ICC, the Inter-American Development Bank's private arm.

(82) After four years of operation, BancoSol is currently *serving* nearly 70,000 clients through twenty-nine offices.

(83) In 1994 the Superintendency of Banks created a new type of regulated financial institution to enable other financially strong non-profit organizations to become regulated and thereby expand the *availability* of financial services to this sector.

In these cases, the verbal complement saturates the theme role. The nominal use of the verb prevents an element of its semantic structure from being expressed (cf. (23)):

(84) *serve* (= (74)):
提供服务
'provide service (to X)' = 'serve Y *to X*'

(85) *be instrumental in* (= (81)):
提供了很大帮助
'provide has very big help (to X) ' = 'help X *with Y*'

The final example exhibits a similar correspondence between the theme of the 'provide' event and the theme (resulting state) of another light verb.

make *available* (= (83):
提供金融服务的可能性
'provide financial service of possibility'

Again the Chinese LCS database allows appropriate English translations to be generated for the verb under consideration. To generate the forms found in the English text, we would again need a systematic way of relating the contents of the complements and the contents of lexicalized roles, which in turn requires a representation for selectional preferences.

5 Conclusions and Future Research

This paper has described how our Chinese LCS database generates and eliminates candidate translations both into and from Chinese based on thematic role structure. The example illustrates (i) how an interlingual thematic representation based in large part on English argument structure may be adapted felicitously to a historically unrelated language, and (ii) how an interlingual (IL) analysis resource can be used in generation. Examination of cases where the candidate translation was represented in the English version of the text show that the refinement to LCS acquisition reported in (Olsen et al., 1998) represents a promising direction for representing cross-linguistic as well as intra-linguistic divergences in predicate argument structure. The model could be refined by a principled representation of selectional preferences, perhaps not directly in the LCS, as well as means of relating the selectional preference of verbs for objects to lexicalized objects.

Acknowledgments

This work has been supported, in part, by DOD Contract MDA904-96-C-1250 and Army Research Laboratory contract DAAL01-97-C-0042. Thanks to members of the following lab groups at Maryland: Computational Linguistics and Information Processing (CLIP), and Language And Media Processing (LAMP), particularly Scott Thomas for his implementation of the grid acquisition program, and John Kovarik, a Chinese language instructor on loan from the Department of Defense.

References

Dorr, B. J., Garman, J., and Weinberg, A. (1995). From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT. *Machine Translation*, 9:71–100.

Dorr, B. J. and Katsova, M. (1998). Lexical Selection for Cross-Language Applications: Combining LCS with WordNet. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, Lanhorne, PA.

Dorr, B. J. and Oard, D. W. (1998). Evaluating resources for query translation in cross-language information retrieval. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain.

Dorr, B. J. and Olsen, M. B. (1996). Multilingual Generation: The Role of Telicity in Lexical Choice and Syntactic Realization. *Machine Translation*, 11(1–3):37–74.

Grimshaw, J. (1993). Semantic Structure and Semantic Content in Lexical Representation. unpublished ms., Rutgers University, New Brunswick, NJ.

Knight, K. and Hatzivassiloglou, V. (1995). Two-Level, Many-Paths Generation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, Cambridge, MA.

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

Levin, B. and Rappaport Hovav, M., editors (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface.* The MIT Press, Cambridge, MA. LI Monograph 26.

Olsen, M. B., Dorr, B. J., and Thomas, S. C. (1998). Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas*, Langhorne, PA.

Otero, M. (1997). América latina: Radiografiá de una proeza. *UNESCO Courier*.

Pinker, S. (1984). *Language Learnability and Language Development*. MIT Press, Cambridge, MA.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. The MIT Press, Cambridge, MA.

Rappaport, M. and Levin, B. (1988). What to do with θ-Roles. In Wilkins, W., editor, *Syntax and Semantics: Vol. 21, Thematic Relations*, pages 7–36. Academic Press, New York.

Resnik, P. (1996). Selectional Constraints: An Information-Theoretic Model and its Computational Realization. *Cognition*, 61:127–159.