Machine Translation, 9:3, 71–100 (1995)

From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT

BONNIE J. DORR bonnie@cs.umd.edu
Department of Computer Science and UMIACS, University of Maryland, College Park, MD 20742

JOSEPH GARMAN garman@umiacs.umd.edu
Department of Linguistics, University of Maryland, College Park, MD 20742

AMY WEINBERG weinberg@umiacs.umd.edu
Department of Linguistics and UMIACS, University of Maryland, College Park, MD 20742

Received September 15, 1994; Revised June 1, 1995

Abstract. Our goal is to construct large-scale lexicons for interlingual MT of English, Arabic, Korean, and Spanish. We describe techniques that predict salient linguistic features of a non-English word using the features of its English gloss (i.e., translation) in a bilingual dictionary. While not exact, owing to inexact glosses and language-to-language variations, these techniques can augment an existing dictionary with reasonable accuracy, thus saving significant time. We have conducted two experiments that demonstrate the value of these techniques. The first tested the feasibility of building a database of thematic grids for over 6500 Arabic verbs based on a mapping between English glosses and the syntactic codes in Longman's Dictionary of Contemporary English (LDOCE) (Procter, 1978). We show that it is more efficient and less error-prone to hand-verify the automatically constructed grids than it would be to build the thematic grids by hand from scratch. The second experiment tested the automatic classification of verbs into a richer semantic typology based on (Levin, 1993), from which we can derive a more refined set of thematic grids. In this second experiment, we show that a brute-force, non-robust technique provides 72% accuracy for semantic classification of LDOCE verbs; we then show that it is possible to approach this yield with a more robust technique based on fine-tuned statistical correlations. We further suggest the possibility of raising this yield by taking into account linguistic factors such as polysemy and positive and negative constraints on the syntax-semantics relation. We conclude that, while human intervention will always be necessary for the construction of a semantic classification from LDOCE, such intervention is significantly minimized as more knowledge about the syntax-semantics relation is introduced.

Keywords: lexical acquisition, interlingual MT, thematic grids, Arabic lexicon, semantic verb classes, syntactic codes, Longman's Dictionary

1. Introduction

Our goal is to construct a large-scale lexicon for each of the languages in an interlingual MT system called PRINCITRAN, a principle-based translator of English, Arabic, Korean, and Spanish (see Dorr (1993); Dorr et al. (1994); Lin et al. (1994)).[1] For many languages, dictionary resources are often scarce, possibly consisting of only a simple bilingual word-list. We describe techniques that predict salient linguistic features of a non-English word using the features of its English gloss (i.e., translation) in a bilingual dictionary. While not exact, owing to inexact glosses and language-to-language variations, these techniques can augment an existing dictionary with reasonable accuracy, thus saving significant time in system development.

As noted by Montemagni and Vanderwende (1992), computational linguistics demands semantic information for parsing, yet little work has been done to extract and use such information on a large scale in a practical application. (See related remarks in (Wilks et al., 1989).) We demonstrate that it is possible to extract enough information from online resources for large-scale, automatic construction of semantic verb classes. The classes allow us to derive thematic role information (for syntactic analysis or lexical disambiguation) and lexical-semantic information (for composition of an interlingua and machine translation). The relevance of this result to machine translation is that translationally equivalent verbs in two languages are likely to share the same collection of thematic roles (i.e., thematic grid) and lexical-semantic representation. The availability of shared thematic and lexical-semantic information in advance facilitates the mapping between a source- and target-language sentence at MT runtime.

We assume that an adequate translation system must include lexical entries for target and source languages that indicate the number of associated syntactic arguments for a particular predicate. We also assume that these arguments can be semantically characterized into particular argument and adjunct types. For example, the thematic grid associated with the verb *hit* might be specified as follows:

agent_theme

These thematic roles are a crucial input to the parsing algorithm, allowing for the association of syntactic constituents with their correct place in a phrase structure tree. They are also crucial to building an interlingual semantic representation—in our case a Lexical Conceptual Structure (LCS) (Dorr, 1993)—in that they provide the syntax-semantics interface between structural positions in the parse tree and the underlying representation of a sentence.

Some have argued that the task of simplifying lexical entries on the basis of broad semantic class membership is complex and, perhaps, infeasible (see, e.g., Boguraev and Briscoe (1989)). However, there have been many arguments for the claim that certain aspects of syntactic behavior are predictable from certain components of meaning, e.g., thematic roles (Fillmore, 1968; Grimshaw, 1990; Gruber, 1965; Jackendoff, 1983; Jackendoff, 1990; Levin, 1993; Pesetsky, 1982; Pinker, 1989). Our working assumption is that thematic-grid information can be correlated with patterns of grammar codes in the Longman's Dictionary of Contemporary English (LDOCE) (Procter, 1978).

While grammar codes have been used previously in automatic extraction tasks (see, e.g., Boguraev and Briscoe (1989); Wilks et al. (1990)), the codes have been used primarily for the prediction of syntactic phrase structure, not for assigning thematic roles that are then used later for semantic analysis. Others who have used the

LDOCE for automatic extraction (e.g., Alshawi (1989); Wilks et al. (1989)) have focused on the derivation of semantic structures from definition analyses rather than the derivation of thematic-grid information from grammar codes. The use of bilingual dictionaries for the automatic construction of multi-language information from LDOCE is also currently under investigation by Farwell, Guthrie, and Wilks (1993). This work is closely related to ours in that it focuses on the extraction of broad semantic restrictions on arguments.[2] The work of Sanfilippo and Poznanski (1992, p. 82) is even more closely related to our approach in that they attempt to create a large lexical database using semi-automatic techniques to recover syntactic and semantic information from machine-readable dictionaries. However, they claim that the semantic classification of verbs based on standard machine-readable dictionaries (e.g., the LDOCE) is "a hopeless pursuit [since] standard dictionaries are simply not equipped to offer this kind of information with consistency and exhaustiveness." Our hypothesis is that a strong correlation exists between some LDOCE code combinations and semantic classes. We expect that the automatic realization of this correlation significantly aids the process of human verification of the resulting lexicon.

The main result of this paper is that our automatic semantic classification of verbs has allowed us build dictionary definitions cross-linguistically. We started with an Arabic-English bilingual dictionary devoid of syntactic and semantic information. (The format of this lexicon is described in Section 2.1.) To use this resource for machine translation on a large scale, we developed efficient acquisition techniques that allow for constructing both structural and semantic information for each entry in the Arabic lexicon. There are a number of benefits to our semantic-classification approach: (1) it provides a framework within which minimal effort is expended to add a new language (e.g., Spanish and Korean) as long as we have access to a bilingual dictionary containing English glosses; (2) it enables automatic acquisition of lexical-semantic representations from which we can construct an interlingua; and (3) it associates with each verb class a set of thematic grids that can be used during syntactic analysis of the input sentence.

We have conducted two experiments that demonstrate the value of our acquisition techniques. The first was designed primarily to demonstrate the feasibility of building a database of thematic grids for over 6500 Arabic verbs; we show that it is more efficient and less error-prone to hand-verify the automatically constructed grids than it would be to build the thematic grids by hand from scratch, even if the initial grids are impoverished.[3] Our approach in this experiment relies on the use of grammar codes from the LDOCE to derive the range of possible thematic-role assignments to verbal arguments. We then make use of a bilingual lexicon to incorporate the resulting thematic grids into an online Arabic dictionary.

The second experiment was designed to demonstrate that it is possible to classify verbs automatically into a richer semantic typology based on Levin (1993), from which we can derive a more refined set of thematic grids. In this second experiment, we show that a brute-force, non-robust technique provides 72% accuracy for semantic classification of LDOCE verbs; we then show that it is possible to approach this yield with a more robust technique based on fine-tuned statistical correlations.

The verb organization resulting from the second experiment provides a more systematic association between verbs and thematic grids than the first experiment: verbs that share a semantic class have the same thematic grid. Moreover, verbal arguments are more accurately specified within each class. In particular, the first experiment used a simple mapping to thematic roles based on the number of arguments together with a very crude semantic classification of those arguments; for example, the subject of a verb was labeled *agent* whether or not this argument was actively causing some action to occur. In the second experiment, the role of the subject could take on additional labels, including *experiencer* or *theme* for cases where the relation between the subject and the action is more passive. We expect that the hand verification of these enriched thematic grids would be at least as efficient as that of the first experiment because the improved grids provide the human checker with much of the information that was missing from the "impoverished" grids. We conclude that, while human intervention will always be necessary for construction of a semantic classification from LDOCE, such intervention is significantly minimized as more knowledge about the syntax-semantics relation is introduced.[4]

The next section describes the first experiment, i.e., the automatic extraction techniques that we applied to the LDOCE in order to incorporate thematic grids into entries in the Arabic lexicon. The results of manual checking for accuracy are also presented. Section 3 describes the second experiment, i.e., our improvements on the automatic acquisition approach used in the first experiment. This experiment involved the use of verb classes from (Levin, 1993) in conjunction with the subcategorization codes in LDOCE. Section 4 discusses the implications of these experiments and examines the possibility of obtaining better results by taking into account linguistic factors such as polysemy and positive and negative constraints on the syntax-semantics relation.

2. Construction of Thematic Grids from LDOCE codes

Our current focus is on the enhancement of the sentence analysis component of PRINCITRAN. As with any sentence analyzer, the lexicon is a significant subcomponent; thus, our primary objective is to construct a lexicon that is rich in both syntactic and semantic information so that we can provide a structural analysis as well as an adequate interlingua from which the target-language sentence will ultimately be generated. We take as our starting point the Alpnet bilingual Arabic-English online dictionary (described below), henceforth referred to as "Alpnet".[5] Our first experiment was designed primarily to test the feasibility of large-scale construction and hand-verification of lexical entries. This involved constructing automatic mappings between English glosses from Alpnet into LDOCE codes; the codes, in turn, were converted into thematic grids which were then exhaustively hand-verified. The experiment demonstrated that it is more efficient and less error-prone to verify the correctness of the resulting grids by hand than it would be to build the thematic grids from scratch, even if the initial grids are impoverished. This section describes the structure of Alpnet and then the methodology and results of the first experiment.

2.1. Structure of the Arabic Lexicon

Alpnet is a broad coverage lexicon containing over 37K lexical entries of which more than 9K are verbs.[6] Because Alpnet was developed as a translation aid for an English speaker, the morphological information for an Arabic word as well as the English gloss were considered sufficient output for a given input word. Given that Alpnet lacked both syntactic and semantic information in its entries, it was critical for us to enhance Alpnet in order for it to become a usable resource as the lexical component in our sentence analysis system. As an example, the syntactic sentence analyzer requires information about argument structures in order to attach potential arguments correctly. Thus, we set about to incorporate thematic grids into the Alpnet lexicon.

Roots in Arabic are significantly unlike roots in Indo-European languages. In Arabic, a root consists of usually 3 but occasionally 4 consonants. There are only about 5000 roots in Arabic. A family of semantically related words is formed by inserting vowels between these consonants as well as attaching prefixes and suffixes. For instance the root *ktb* gives rise to a large number of words involving ‘writing’. Some of the entries for *ktb* are shown in Figure 1.

Arabic	**Gloss**
k_t_b_(Iau,high)	write
k_t_b_(III)	correspond with
k_t_b_(IV)	dictate; make write
k_t_b_(VI,intr)	correspond
k_t_b_(VII)	subscribe
k_t_b_(VIII)	register,enroll
k_t_b_(X)	make write; dictate
k_t_b_(iA-Ndu,high)	book
k_t_b_(uu-N,high)	books
k_t_b_(maoaiy∼-N-ap,high)	office
k_t_b_(aaap-Nap)	authors; writers

Figure 1. Entries for Arabic Root *ktb*

The lexical form of the root *ktb* is encoded in the dictionary by the string k_t_b_, which indicates the consonants used to realize various forms of the root. Each entry indicates which vowels to use to form a particular word, the part of speech, and possibly other information. For example, the entry ‘Iau,high’ encodes certain information concerning the possible surface realizations of the root form k_t_b_:

'I'	the word is a form I verb (forms are enumerated, I-X)
'au'	the perfect stem is *katab* and the imperfect stem is *ktub*
'high'	the word occurs with high frequency

The format given above is for open class words, e.g., nouns, verbs and adjectives. The Alpnet lexicon contains 4914 roots with 37355 root entries (words). This corresponds to essentially all the roots in Arabic. Of the 37355 entries in the Alpnet lexicon, 9460 are verbs. Most verbs have more than 1 gloss (e.g., see the third entry with the gloss 'make write/dictate' for *ktb* given above). The total number of glosses for 9460 verbs is 17971. Without duplicates, there are a total of 6587 unique glosses. This information is summarized in Table 1.

Table 1. Number of Arabic Verbs and English Glosses

Type	**# Occurrences**
Arabic verb entries	9460
English verb glosses	17971
unique English verb glosses	6587

Table 2 shows how the 17971 English glosses were distributed over the 9460 Arabic verbs. In particular, there is one occurrence of an Arabic verb with 6 English glosses, 10 occurrences of an Arabic verb with 5 English glosses, and so on.

Table 2. Number of English Glosses Per Occurrence of Arabic Verb

# Glosses	**# Occurrences**
1	2832
2	4987
3	1411
4	219
5	10
6	1

Some sample glosses are: *Americanize, Anglicize, Westernize, abandon, abate, abstain from, be aggressive, be free, consider faithless, rub the skin with liniment, wink to each other, act cautiously, be mobilized, flow continuously,* and *writhe seductively*. Note that the glosses may be single words or multi-word phrases.

2.2. Assignment of LDOCE-Based Thematic Grids to Arabic Lexicon Entries

The task of assigning thematic grids to entries in the Arabic lexicon consists of two steps: (1) automatic mapping of English glosses to precompiled thematic grids from an English dictionary; and (2) manual checking of these English-based grids as valid grids in their corresponding Arabic verb entry. The manual checking was carried out by a native Arabic speaker. We present the technique of thematic grid compilation from the English dictionary and then describe the mapping from the English glosses to the compiled thematic grids. We also discuss the accuracy and efficiency of this approach as determined through hand-verification of the resulting thematic grids.

LDOCE contains extensive information about argument structure for English verbs. Each verb in this dictionary is coded for the arguments that it can take. We have used an on-line version of the LDOCE to extract verbs along with their syntactic "codes" (i.e., subcategorization frames). We decode these into a corresponding set of thematic grids for each verb entry according to the mapping given in Table 3.[7] (The following abbreviations are used in this figure: ag = agent; th = theme; adv = adverb; pred = predicate; ben = benefactive; inf = infinitive; prop-subj = embedded complement with no subject; prop+subj = embedded infinitive with subject; inf = infinitive; prop+that = that-headed complement; prop+wh = wh-headed complement.) We use the following conventions to specify thematic grids: (1) The first thematic role in the grid is obligatory; (2) Any thematic role preceded by an underscore (_) is obligatory; (3) Any thematic role preceded by a comma (,) is optional; (4) Any thematic role that is followed by a parenthesized preposition must necessarily be contained in a phrase headed by that preposition.[8] Although conventions (3) and (4) do not apply to the grids in Table 3, they are relevant to the enhanced thematic grids described in Section 3.

By using the English glosses in the Arabic dictionary and these thematic grids derived from the LDOCE, we were able to install thematic grids automatically into all Arabic verb entries. After looking up the grid for each English gloss of an Arabic verb, a consensus heuristic was applied. When the verb had only 1 gloss, the thematic grid of that gloss was taken. However, when the verb had 2 or more glosses, a thematic grid (other than 'ag' or 'ag_th') must occur at least twice to be accepted. The grids 'ag' and 'ag_th' need only occur once to be accepted. The lower criterion for these grids makes sense when one realizes that these forms basically stand proxy for transitive and intransitive verbs. The other codes are finer distinctions on these basic categories and so in some sense count as occurrences of these categories. This heuristic ruled out many idiosyncratic thematic grids while maintaining the common thematic grids. The resulting thematic grids were then written into the lexicon.

All 9460 Arabic verb entries had thematic grids added automatically. Out of the 6587 unique English glosses, 2696 exactly matched a verb entry in the LDOCE. In these cases, our program directly incorporated the thematic grid from the code as given in Table 3. These glosses were typically single verbs or single verbs followed by a preposition. Some examples are shown in Table 4.[9]

Of the remaining glosses, 2721 were what we call "parsed matches", i.e., matching phrases that contain a verb followed by a noun phrase, prepositional phrase, preposition or some combination thereof. Phrases ending with a preposition were considered transitive (T1), e.g., *abstain from*. Phrases ending with a noun phrase were considered intransitive (I), e.g., *adorn oneself*.[10] Phrases ending with a prepositional phrase were handled according to the preposition: verbs associated with

Table 3. LDOCE Codes, Thematic Grids, and Examples

Code	Thematic Grid	Example
I	ag	we **paused**
I	th	the water **froze**
L9	ag_adv	she **lives** here
L1	ag_pred	she **became** queen
L7	ag_pred	she **became** famous
T1	ag_th	she **kicked** the boy
X9	ag_th_goal	**put** it in the box
D1	ag_ben_th	**give** the boy a book
X1	ag_th_pred	she **considered** him her enemy
X7	ag_th_pred	she **considered** him dead
I2	modal	I **can** fly
T2	ag_event+inf	I **helped** clean the window
I3	ag_prop-subj	he **lived** to be 90
L3	th_prop-subj	the difficulty **is** to know what to do
T3	ag_prop-subj	I **want** to go
V2	ag_event+subj	he **saw** her leave
V3	ag_prop+subj	he **wants** her to leave
I5	prop+that	it **appears** that she will win
I6	prop+wh	it **appears** as if she will win
L5	th_prop+that	trouble **is** that you know bill
L6	th_prop+wh	it **is** as if we had never met
T5	ag_prop+that	I **know** that he will come
T6	ag_prop+wh	he **decided** who should go
D5	ag_ben_prop+that	he **warned** her that he runs
D6	ag_ben_prop+wh	**tell** me who is here
I4	ag_event+ing	she **came** running
L4	ag_event+ing	she **ended up** dancing
T4	ag_event+ing	I **enjoyed** singing
V4	ag_th_event+ing	he **watched** her cooking dinner
L8	th_event+ed	he **got** trapped
I8	sj_event+ed	smoking **is** not permitted
V8	ag_th_event+ed	he **had** a house built

Table 4. English Glosses that Match an LDOCE Verb Entry Exactly

Arabic	English Gloss	LDOCE Code	Thematic Grid(s)
'_n_kl_z_(QI)	Anglicize	T1	ag_th
x_*_l_(Iau)	abandon	T1	ag_th
b_x_w_(Iau,intr,low)	abate	I T1	ag ag_th
H_z_b_(III)	adhere to	T1	ag_th
y_m_X_(V)	aim at	T1 T4	ag_event+ing
k_w_f_(V,intr)	band together	I	ag

the prepositions *of*, *as*, or *by* were considered both transitive and intransitive (I T1); verbs associated with other prepositions were considered intransitive. Because *to* can be both a preposition and an infinitive marker, phrases that contained *to* were checked by hand. Some examples of parsed matches are shown in Table 5.

Table 5. English Glosses that Correspond to a Verb Phrase Construction

Arabic	**English Gloss**	**LDOCE Code**	**Thematic Grid(s)**
f_T_m_(VII)	abstain from	T1	ag_th
m_$_y_(III,med-high)	act in unison with	T1	ag_th
'_l_f_(III)	adapt to	T1	ag_th
m_$_H_(Iaa,low)	administer extreme unction to	T1	ag_th
h_y_'_(VI)	accommodate each other	I	ag
f_h_m_(VI)	achieve mutual understanding	I	ag
j_m_l_(V,intr)	adorn oneself	I	ag
m_w_h_(II)	add water	I	ag
k_f_r_(IV)	accuse of infidelity	I T1	ag ag_th
s_h_d_(II)	deprive of sleep	I T1	ag ag_th
r_$_H_(II)	appoint as candidate	I T1	ag ag_th
r_h_n_(X)	demand as a security	I T1	ag ag_th
s_b_E_(II)	divide by seven	I T1	ag ag_th
E_n_q_(II)	grab by the collar	I T1	ag ag_th
x_l_S_(VI)	act with integrity	I	ag
x_l_S_(VI,intr)	be at a distance	I	ag
w_s_T_(V)	be in the middle	I	ag
s_q_y_(X)	pray for rain	I	ag
n_q_d_(III)	call to account	I T1	ag ag_th
'_*_n_(II,low)	call to prayer	I T1	ag ag_th
n_t_n_(II)	cause to decay	I T1	ag ag_th
w_q_f_(V,act,high)	come to a halt	I	ag
H_S_H_S_(QI)	come to light	I	ag

An additional 642 glosses were classified as "-ed past participles". These phrases consist of a verb followed by a '-ed' past particle. The past participles were considered to be adjectives and, thus, consumed one thematic role of the head verb. Generally this made the phrase intransitive (I), except in the case of *make*, which was both transitive (T1) and intransitive (I). Some examples of "-ed past participles" are: *be Anglicized*, *be aroused*, *become Africanized*, *feel embarrassed*, *get burned*, and *make bored*.

There were 116 glosses that fell into the category of "-ly adverbs"; these consist of a verb followed by a '-ly' adverb. In such cases, the code of the head verb was used.[11] Some examples are given in Table 6.

An additional 118 glosses were classified as '-s plurals'. These are phrases ending with plural nouns. The rules which apply to the "parsed matches" were applied here. Some examples are *appear in the heavens*, *apply the brakes*, *be concerned with trifles*, *be covered with warts*, *flow in torrents*, *smack the lips*, and *use stratagems*.

There were 231 remaining phrases with "missing" words. These were phrases containing words not found in the LDOCE. Frequently, these phrases contained spelling errors, derived words, foreign words, or unusual words. A small number

Table 6. English Glosses that Match an LDOCE Verb Plus an -ly Form

Arabic	English Gloss	LDOCE Code	Thematic Grid(s)
H_*_r_(V)	act cautiously	I L1 L9 T1	ag ag_pred ag_adv ag_th
d_r_j_(V)	advance gradually	I T1	ag ag_th
d_n_y_(V,intr)	approach gradually	I T1	ag ag_th
w_l_y_(VI)	arrive constantly	I	ag
w_q_Z_(Iai,low)	beat brutally	I L9 T1 X9	ag ag_adv ag_th ag_th_goal
s_d_r_(Iia)	behave indifferently	I L9 T1	ag ag_adv ag_th
q_f_l_(IV)	close securely	I T1	ag ag_th
E_n_f_(IV)	deal with harshly	D1 I T1	ag_th ag_ben_th
n_q_S_(VI)	decrease gradually	I T1	ag ag_th
E_b_T_(Iai)	die prematurely	I L1 L7 L9 T1	ag ag_pred ag_adv ag_th
n_h_d_(VI)	distribute equitably	T1	ag_th
E_m_d_(Iai,intr)	do deliberately	D1 I I2 L7 L9 T1	ag_adv ag_th ag_ben_th ag_pred modal
q_t_X_(Iau)	render falsely	D1 T1 X7	ag_th ag_ben_th ag_th_pred
$_d_q_(V)	speak affectedly	I L9 T1	ag ag_adv ag_th
h_l_h_l_(QI)	weave flimsily	I L9 T1 X9	ag ag_adv ag_th ag_th_goal
n_q_D_(IV)	weigh heavily	L1 L9 T1 X9	ag_pred ag_adv ag_th ag_th_goal
n_$_T_(V,intr)	work energetically	I L9 T1 X9	ag ag_adv ag_th ag_th_goal
l_f_X_(II,med-low)	wrap tightly	T1 X9	v ag_th ag_th_goal
q_S_E_(V)	writhe seductively	I	ag

were legitimate words that were not available in the LDOCE. Verbs in this category were considered optionally transitive or intransitive (T1 I).[12] Some examples are: *Africanize*, *Arabicize*, *Bolshevize*, *be an extremist*, *be blackish*, *be boring*, *acclimate*, *acquaint*, *administrate*, *adsorb*, *dance the dabka*, *deceive*, *demobilize*, *increase tenfold*, and *say "phew"*.

Finally, there were 63 words found in the LDOCE that could not be assigned a thematic grid, i.e., the LDOCE had no codes for these verbs. In these cases, the verbs were considered transitive or intransitive (I T1). Some examples are: *advert*, *armor*, and *attribute*.

The distribution of verb glosses resulting from the above analysis is summarized in Table 7. All 9460 verb entries had thematic grids added automatically using

Table 7. LDOCE Code Assignments for 6587 Verb Glosses

Type of Match	# Glosses	Code(s) Assignment	Example
Exact match	2696	LDOCE code	see Table 4
Parsed match	2721	LDOCE Code for V	see Table 5
-ed Past Participle	642	I or T1 I (for *make*)	s_'_m_(IV) (make bored)
-ly Adverb	116	LDOCE Code for V	see Table 6
-s Plural	118	I	H_w_l_(VIII) (use stratagems)
Not found	231	I T1	m_l_X_(IV,intr) (be boring)
Found, but no code	63	I T1	$_E_r_(IV) (advert)

the mapping provided in Table 3. We show, as an example of our results, the augmented entries for the root *ktb* in Figure 2.

Arabic	**Gloss**	**Thematic Grids**
k_t_b_(Iau,high)	write	ag ag_th ag_ben_th ag_prop-subj ag_prop+that ag_ben_prop+that ag_event+ing
k_t_b_(III)	correspond with	ag_th
k_t_b_(IV)	dictate; make write	ag ag_th
k_t_b_(VI,intr)	correspond	ag
k_t_b_(VII)	subscribe	ag ag_th ag_th_pred
k_t_b_(VIII)	register,enroll	ag ag_th
k_t_b_(X)	make write; dictate	ag ag_th

Figure 2. Entries for *ktb* after Thematic-Grid Installation

2.3. Verification of Correctness

The results described in the last section were checked by a native Arabic speaker. Table 8 lists each LDOCE-derived thematic grid along with the number of English glosses that were automatically assigned a grid (**Attempts**) and the number of Arabic verbs (corresponding to those English glosses) for which the grid was attested

Table 8. Verification of Automatic Acquisition by Native Arabic Speaker

Thematic Grid	Attempts	Successful
ag	7840	7840
ag_adv	1377	380
ag_pred	362	74
ag_th	6397	6397
ag_th_goal	1172	317
ag_ben_th	609	142
ag_th_pred	485	93
modal	18	2
ag_event+inf	4	0
ag_prop-subj	631	152
th_prop-subj	13	2
ag_event+subj	45	8
ag_prop+subj	598	163
prop+that	40	10
prop+wh	29	9
ag_prop+that	770	219
ag_prop+wh	323	83
ag_ben_prop+that	68	26
ag_ben_prop+wh	24	8
ag_event+ing	490	114
ag_th_event+ing	162	22
th_event+ed	0	0
sj_event+ed	3	2
ag_th_event+ed	13	2

by the Arabic speaker (**Successful**).[13] Approximately 90% of the thematic grids that were automatically added were attested by the native speaker. Following the heuristics given above, the numbers are the same for 'ag' and 'ag_th'. Recall that only one gloss is needed in order to attest these thematic grids, while for the other thematic grids, at least two glosses must have the thematic grid in order for it to be attested.

This experiment was designed to demonstrate the feasibility of building thematic grids on a large scale in a given language. Our automatic acquisition program ran in only a few minutes on the entire set of Alpnet verbs and the lexical verification process took only two weeks by a single native Arabic speaker. We estimate that it would take at least 6 months for a native speaker to build such a lexicon from scratch (by human recall and data entry alone), and in such a case, the potential for error would be at least twice as high. We base this estimate on: (1) a manual construction time of 5 minutes per entry (given all the possible usages of a verb) for 9460 verb entries, totaling 5 months, plus (2) an additional month of proofing and consistency checking across entries. Note that, because the LDOCE entries are already consistency-checked (through years of revisions by teams of lexicographers), the error rate of the LDOCE-derived Arabic lexicon is significantly lower than that of a lexicon constructed from scratch by a single Arabic speaker, unless a heavier person-hour commitment can be made to the proofing process.[14]

Clearly, the thematic grids that have been acquired in this experiment provide minimal thematic relations based on simplistic subcategorization frames (e.g., transitive and intransitive). While subcategorization frames are useful for syntactic processing, they do not describe the full range of thematic possibilities that would be necessary for construction of an interlingua for machine translation. This was the motivation for moving toward more enriched thematic roles, as developed in the second experiment described in the next section.

3. Construction of Thematic Grids from Levin's Classification

While the information in the LDOCE has aided us significantly in our endeavor to construct a large-scale lexicon containing thematic grids, we are still left with two problem: many of these grids are highly language specific (e.g., 'th_event+ed', which has no analog in Arabic as indicated in Table 8) and many thematic roles have not been included (e.g., source, location, time, etc.). Thus, our next step was to define thematic grids on the basis of a finer-grained verbal classification along the lines of (Levin, 1993). This approach has a better chance of capturing subtle meaning distinctions while still supporting an interlingual MT framework.

We have built an alternative set of thematic grids that are richer than the ones described above as well as more language-independent. A small sample of these is shown in Table 9.[15] By richer, we mean that argument positions are associated with a wider variety of semantic roles, e.g., intransitives are not uniformly marked 'ag', but may be marked 'th', depending on the real logical role of the argument. For example, the semantic class 29.2 in (Levin, 1993) ("Characterize Verbs") is associated with the thematic grid 'exp_perc_instr(as)'.[16] By language-independent, we refer to more syntactically neutral thematic specifications such as 'exp' (experiencer) and 'perc' (perceived entity) rather than 'event+ed'. We pursue Levin's hypothesis that semantically related verbs have similar, or identical thematic grids, i.e., we assume that verbs of the same semantic class have the same thematic specification. Thus, all verbs in class 29.2 (e.g., *accept, remember, employ,* and *visualize*) share the grid 'exp_perc_instr(as)'. We also assume that these thematic grids apply to the verb semantics of other languages and that certain syntactic features (e.g., phrasal type and prepositions) can be readily factored out and refined by the native speaker.

Table 9. Thematic Grids Based on Levin's Verb Classification

Levin Verb	**Levin Class**	**Thematic Grid**
admire	31.2 "Admire"	exp_perc,purp(for),instr(as)
act	29.6 "Masquerade"	th_pred
buy	13.5.1 "Get"	ag_th,src(from),poss(for),ben(for)
hold	15.1 "Hold"	ag_th,instr(by)
keep	15.2 "Keep"	ag_th,loc
mix	22.1/36.1 "Mix/Correspond"	ag_th,manner(together) ag_th,goal(with)
pronounce	29.3 "Dub"	ag_th_pred
put	9.1 "Put"	ag_th_goal()
remember	29.2 "Characterize"	exp_perc_instr(as)
remove	10.1 "Remove"	ag_th,src
tear	23.2/45.1 "Split/Break"	ag_th_src(off of,off) th

We have manually encoded thematic grids for all 192 of Levin's semantic classes.[17] These classes consist of a total of 2775 unique verbs distributed across 3828 Levin entries, where an *entry* is a "class/verb" pair, e.g., "29.2/accept". Our intent is to "port" these hand-coded thematic grids into the Alpnet dictionary using the same technique described above, thus replacing the LDOCE-based thematic grids with the richer version. Of the 6587 unique English glosses in the Arabic dictionary, we found 3298 in (Levin, 1993); thus, we achieved about 50% overlap with the verbs in the entire Arabic dictionary.[18]

Of the 3289 missing glosses, it is interesting to note that 1710 are of the form "be X," 161 are of the form "become X," 65 are of the form "have X," and 1353 are of some other form. As seen by these numbers and the regular structure of these gloss patterns, the mapping from Alpnet glosses to Levin's verbs is a potentially productive process. Random examples of verbs not found in (Levin, 1993) are shown in Figure 3.

This section describes an experiment that will ultimately lead to the incorporation of an enriched set of thematic grids into the Alpnet lexicon using a combination of LDOCE codes and the verb classes in (Levin, 1993). The experiment presented below tests the hypothesis that there exists a correspondence between code combinations from the LDOCE and Levin's semantic classes. While our initial attempts to map *individual* subcategorization codes to thematic grids (in the previous experiment) yielded an impoverished set of grids, we report here that better results for Alpnet thematic-grid assignment can be achieved by mapping *combinations* of

Multi-Word Verbs (be): be a coward, be aware, be colorful, be delivered, be engrossed, be gathered, be in the sun, be lame, be newly acquired, be possible, be separated from, be suitable, be valid

Multi-Word Verbs (become): become a Christian, become difficult, become numerous, become thin

Multi-Word Verbs (have): have a bad character, have a cold, have no market, have pity, have weak eyesight

Other: adjust, apprehend, approach, approach gradually, bear besiege, compel, deceive, distance, fail, harbor, harbor resentment, immunize, lack, mutilate, ponder, ponder ways and means, reflect, secede, summon, sunbathe, venture, wrong

Figure 3. Sample Alpnet verbs not found in Levin

subcategorization codes (henceforth *code patterns*) into Levin's semantic classes, each of which is associated with a finer-grained thematic grid.

We show that the presence of certain code patterns allowed us to determine, automatically, the semantic classification of 72% of all verbs in (Levin, 1993) using a brute-force, non-robust technique. We then demonstrate that it is possible to approach this yield with a more robust technique based on fine-tuned statistical correlations.

3.1. Brute-Force Semantic Classification of Arabic Entries

Our first task in this experiment was to determine how well we can classify LDOCE verbs using a brute-force (hence non-robust) technique for semantic classification based on a correlation between unique LDOCE code patterns and Levin's semantic classes. We started by examining LDOCE-coded verbs that occur in (Levin, 1993). For each such verb, we determined the probability that the verb (with its associated LDOCE codes) belonged to a given Levin-based class. We called this probability $L(\text{class}|\text{codes})$, i.e., the probability of predicting a Levin-based class given a LDOCE code pattern, where the Levin-based class with the highest probability is the best choice, the next highest probability is the next best choice, etc. The objective was to set $L(\text{class}|\text{codes})$ for every class-codes combination so that the resulting classification would be as close as possible to the "real" classification (i.e., the classification provided by Levin). No regard was given to anything else other than achieving a maximally correct verb classification with respect to (Levin, 1993) using the LDOCE coding scheme. In particular, we did not attempt to take different word senses into account, nor did we use linguistic constraints that might be derived from a closer examination of the relation between syntactic codes and semantic classes. These points will be discussed briefly in Section 4 and are addressed more thoroughly in (Dorr and Jones, 1995).

The optimal value for each $L(\text{class}|\text{codes})$ was defined on the basis of a frequency count on occurrences of LDOCE-based code patterns (abbreviated "codes") in the Levin-based classes (abbreviated "class"). We first constructed a list of all unique LDOCE-based code patterns and then determined the number of times each pattern appeared on entries in each Levin-based semantic class. There were 925 unique LDOCE-based code patterns, over half (464) of which occur only once, and 849 of which occur 5 times or less. There were 2775 unique verbs representing 3828 entries in Levin's classes (some verbs occur in multiple classes).

For a given semantic class and code pattern, we defined $L(\text{class}|\text{codes})$ to be equal to the number of times that the code pattern appeared in that class, divided by the number of times that code pattern appeared in all classes. In order to fine-tune the statistical correlations, we used an enhanced set of LDOCE codes. These codes are as discussed above with the addition of 'N' for a verb that is associated with a nominal form (e.g., *kick*) and 'ADJ' for a verb that is associated with an adjectival form (e.g., *free*); also, certain categories are enhanced with prepositional selections (e.g., T1-WITH for the verb *illuminate*) which figure heavily in Levin's semantic classification. This enhanced LDOCE coding system, which we derived with a fully automatic procedure, allows us to map more precisely from the LDOCE entries into Levin's semantic classes. Table 10 provides a very small sample of the enhanced codes (37 out of 174) with their corresponding argument structures and example of verbs from (Levin, 1993); for the expanded version of this table, see (Dorr and Jones, 1995).

Table 11 shows the result of this experiment for 31 out of 925 unique LDOCE code patterns (under the column labeled **Pattern**). The top 5 Levin-based semantic classes for each pattern are given under the column labeled **Levin Classes**. The parenthesized numbers correspond to Levin's semantic classes and the unparenthesized numbers represent the number of occurrences of the codes associated with these classes. For example, the LDOCE code pattern 'I T1 N' appeared 21 times with verbs from Levin's class 37.3 ("Manner of Speaking Verbs") which contains, for example, the verbs *babble* and *bark*.[19] The column labeled **All** corresponds to the total number of occurrences of a LDOCE-based code pattern in the full set of 3828 Levin entries; the column labeled **Top 5** corresponds to the total number of occurrences in the top 5 semantic classes.

The probability $L(\text{class}|\text{codes})$, as defined, guarantees a classification that is optimal, assuming our only inputs to the procedure are Levin's semantic classification and the enhanced code combinations from LDOCE. To show this, we consider an example. As shown in Table 11, the LDOCE-based code pattern 'I T1 N' appeared 265 times. This pattern occurred 21 times in the semantic class 37.3, the most common class for this code pattern. Therefore, $L(\text{class} = 37.3|\text{codes} = \text{'I T1 N'}) = 21/265$. When we attempt to classify the pattern 'I T1 N', $L(\text{class}|\text{'I T1 N'})$ will equal the number of times 'I T1 N' occurred divided by 265. The class with highest probability is simply the class with the highest occurrence, (e.g., class 37.3 in the

Table 10. Enhanced LDOCE Coding System Associated with Levin's Verbs

LDOCE code	**Arguments**	**Adjuncts**	**Example**
D1	NP NP	–	allow
D1-FOR	NP NP or NP PP[for]	–	build
D1-OVER	NP PP[over]	–	scatter
D1-TO	NP NP or NP PP[to]	–	deliver
D1-WITH	NP PP[with]	–	entrust
D5	NP CP[that+fin]	–	grant
I	–	–	doze
I-ABOUT	–	PP[about]	learn
I-WITH	–	PP[with]	glisten
I3	VP[to+inf]	–	yearn
I4	VP[+prog]	–	go
I5	CP[that+fin]	–	wail
L1	NP	–	remain
L7	AP	–	keep
L8	VP[+past-part]	–	get
L9	ADV/PP	–	live
L9-WITH	PP[with]	–	settle
T1	NP	–	kick
T1-FROM	NP	PP[from]	excuse
T1-OF	NP	PP[of]	clear
T1-TO	NP	PP[to]	escort
T1-WITH	NP	PP[with]	wad
T2	VP[+inf]	–	help
T3	VP[to+inf]	–	want
V2	NP VP[+inf]	–	make
V3	NP VP[to+inf]	–	design
V4	NP VP[+prog]	–	hear
V8	NP VP[+past-part]	–	need
WA5	-ed adjectival	–	scramble
WV2	VP[+inf]	–	can
WV3	-ing morphology	–	quibble
WV4	-ing adjectival	–	grunt
WV5	-ed adjectival	–	scramble
X1	NP NP	–	consider
X1-TO_BE	NP VP[to+be+NP]	–	acknowledge
X5A	CP[(that)+fin]	–	warrant
X9	NP ADV/PP	–	swab

Table 11. Sample of Unique LDOCE Code Patterns and with their Frequencies in Top 5 Levin-based Semantic Classes

LDOCE Code Patterns	**Levin Classes**	**All**	**Top 5**
I T1 N	21 (37.3) 16 (45.4) 112 (43.2) 11 (38) 9 (22.4)	265	69
T1 N	31 (9.9) 20 (31.1) 15 (22.4) 13 (9.8) 13 (9.10)	246	92
T1	52 (31.1) 16 (33) 14 (45.4) 10 (44) 9 (10.5)	220	101
I T1	85 (45.4) 6 (47.2) 5 (43.2) 5 (48.1.1) 4 (10.5)	188	105
N	20 (9.9) 15 (51.4.1) 12 (13.7) 9 (10.7) 9 (51.5)	145	65
I N	18 (38) 12 (37.3) 9 (36.1) 9 (43.2) 8 (40.2)	120	56
I	5 (51.3.2) 4 (48.1.1) 3 (45.4) 3 (45.5) 3 (48.2)	56	18
I T1 X9 N	4 (22.4) 3 (10.4.1) 3 (11.4) 2 (10.7) 2 (17.1)	52	14
I L9 T1 X9 N	5 (51.3.2) 3 (22.4) 3 (47.3) 2 (17.1) 2 (18.4)	47	15
L9 T1 X9 N	3 (9.3) 2 (10.4.1) 2 (17.1) 2 (18.1) 2 (22.4)	32	11
L9 N	9 (51.3.2) 4 (56) 2 (30.3) 2 (43.2) 2 (47.6)	28	19
L9	13 (51.3.2) 4 (47.1) 2 (46) 1 (31.3) 1 (36.1)	27	21
T1 T1-WITH N	7 (9.9) 6 (9.8) 2 (31.1) 2 (47.8) 1 (9.3)	23	18
T1 WV4	17 (31.1) 2 (39.4) 1 (27) 1 (33) 1 (47.8)	22	22
T1 T1-FROM	7 (10.1) 3 (10.5) 2 (10.4.1) 2 (10.6) 2 (23.1)	20	16
I WV4 N	2 (40.2) 2 (43.1) 2 (47.2) 2 (47.6) 1 (31.3)	19	9
X9 N	4 (22.4) 2 (29.8) 1 (9.1) 1 (9.7) 1 (10.4.1)	17	9
L9 X9 N	2 (51.3.2) 1 (9.2) 1 (11.5) 1 (19) 1 (22.3)	16	6
I I-WITH T1 N	5 (43.2) 2 (36.1) 1 (37.3) 1 (38) 1 (40.2)	14	10
I L9 T1	2 (45.4) 1 (9.5) 1 (9.7) 1 (22.3) 1 (26.1)	13	6
I I-ABOUT N	2 (13.7) 2 (37.3) 2 (37.6) 2 (38) 1 (37.8)	12	9
I I-FOR T1 N	2 (35.1) 2 (35.2) 1 (10.9) 1 (13.7) 1 (31.3)	11	7
T1 T1-OF	9 (10.6) 1 (45.4) 0 () 0 () 0 ()	10	10
D1 D1-FOR I L9 T1 X7 X9 N	1 (9.6) 1 (11.2) 1 (22.3) 1 (23.2) 1 (26.1)	9	5
I T1 T1-WITH N	2 (22.2) 1 (9.7) 1 (9.8) 1 (36.1) 1 (43.4)	8	6
I L9 T1 T1-ON WV5 X9 N	1 (10.1) 1 (12) 1 (23.2) 1 (25.2) 0 ()	4	4
I L9 T1 T1-WITH N	1 (17.2) 1 (51.3.2) 1 (57) 0 () 0 ()	3	3
D1-ON D1-OVER D1-WITH I T1 N	2 (9.7) 0 () 0 () 0 () 0 ()	2	2
T1 WV5 WV5-WITH N	1 (9.9) 0 () 0 () 0 () 0 ()	1	1
T1 T1-IN T1-WITH X7 N	1 (9.9) 0 () 0 () 0 () 0 ()	1	1
T1 T1-IN T1-WITH WV5 N	1 (9.9) 0 () 0 () 0 () 0 ()	1	1

example above). In this way we have correctly classified the maximum number of verbs. Put another way: if any other class were selected as the best first guess for a particular code pattern, fewer verbs would have been correctly classified with that code pattern. Likewise, the best second choice is also chosen, and so on.

3.2. Results of Brute-Force Classification Strategy

We evaluated the classification strategy given above on the 3828 Levin entries. (The intent, then, is to compare the results with the actual classification given in (Levin, 1993).) The number of entries correctly classified is shown in Table 12. Of the 3828 entries, 2659 (69.5%) were correctly classified in one of the 5 best classes. Recall that, because Levin's classification lists some verbs in more than one entry, the 3828 entries actually correspond to only 2775 unique verbs. Of these 2775 verbs, 2009 (72.4%) had at least one class correctly chosen.

Table 12. Semantic Classification Based on Brute-Force Strategy

Choice					Total	Total
1st	2nd	3rd	4th	5th	(out of 3828 entries)	(out of 2775 verbs)
1402	587	331	200	139	2659 (69.5%)	2009 (72.4%)

As described in the last section, the probability L was defined such that the maximum number of Levin verbs would be classified correctly. That is, for each class/code combination, it is not possible to set L any other way without lowering the number of correct classifications for a given code. While the statistical method is a brute-force technique, it does serve to show that the upper bound for *any* unconstrained classification scheme based on statistical correlations between LDOCE code patterns (without sense disambiguation) and Levin's semantic classes is ~72%. That is, we will never do better than this, as long as we ignore different word senses and linguistically motivated constraints that relate the codes to semantic classes.

While it is possible to address the problem of word sense disambiguation and imposing linguistic constraints (see Section 4), a more serious difficulty is that the brute-force technique is not robust. If we were to achieve 72% accuracy consistently using this approach, our efforts to build an Arabic dictionary would receive a considerable boost in that the human checker would only have to make changes less than 30% of the time. However, we must consider whether the method is useful for the acquisition of novel verbs, i.e., those *not* occurring in (Levin, 1993). In particular, as mentioned in the previous section, over half (464) of the 925 unique LDOCE patterns occurred only once. One might wonder how many additional unique LDOCE patterns exist (outside of the ones corresponding to Levin's verbs) and what we should do when such patterns are encountered. The current classification scheme also does not handle new LDOCE patterns. An analysis of the remaining verbs in the LDOCE indicates, in fact, that there are over 600 additional unique codes (a total of 1524 unique codes). Thus, the ability to handle new codes is clearly an important issue that we must address. (We will return to this point in the next subsection.)

Note that, even if the number of new LDOCE patterns is low, the fact that half of the 925 unique LDOCE patterns occur only once indicates that the 72% statistic is spuriously high; obviously, it is not difficult to guess the semantic class of a verb whose LDOCE code occurs only once—the class is, in fact, always correctly determined in these cases. This is an inherent danger of *any* analysis that assumes a predictable input (in this case, a perfect code-to-class correspondence even for verbs whose code is unique). When we consider the assumptions that were made to obtain the above results, it is not difficult to see how we achieved such a high success rate of 72%. The next section describes an experiment that addresses the issue of robustness.

3.3. Robust Semantic Classification of Arabic Entries

As we have already seen, we will never be able to surpass 72% accuracy using brute-force classification based on monolithic code patterns from LDOCE. For example, the verb *sleep* (as in appears in *Gloria slept*) appears in Levin's class 40.4 ("Snooze Verbs") as well as class 54.3 ("Fit Verbs") (as in *The cabin sleeps 5 people*). Both of these senses of *sleep* appear in the LDOCE (the first with the code 'I' and the second with the code 'T1'), yet our current algorithm collapses these two into the same code pattern 'I T1'. Moreover, the algorithm is unable to determine whether there is any connection between such verbs and other verbs that have this code pattern as a sub-pattern (e.g., *nap*, which also occurs in class 40.4 but has the pattern 'I T1 N'). In an effort to overcome this difficulty, we examined correlations between the individual codes that make up the pattern and the semantic classes of Levin. This investigation has led to the development of a robust technique that applies to new LDOCE code patterns as they are encountered. With this approach we were able to achieve 55% accuracy of semantic classification (automatically checked against the verbs occurring in (Levin, 1993)).

The approach is not to analyze LDOCE codes in terms of a correspondence between code patterns and semantic classes, but rather as a correspondence between individual codes and semantic classes. Note that this is different from the experiment reported in Section 2 which maps individual codes directly to (impoverished) thematic grids without regard to semantic classification. We believe this to be the first step toward the application of linguistic techniques for building a Levin-based semantic classification. In particular, the results of this experiment have revealed a correlation between syntactic combinations of behaviors and LDOCE codes that might be exploited in future experiments.

We first built a cross-reference table of Levin's classes with LDOCE codes. For each Levin entry, we associated a set of LDOCE codes with the "class/verb" pair, resulting in a table of triplets of the form "class/verb/codes". (Levin entries that do not occur in LDOCE were not considered.) Next we developed a correlation function (CORR) for matching a given code to a given class. The probability that a code is associated with any given class is the total number of occurrences of that code in the cross-reference table, divided by the total number of triplets in the table. Therefore, the number of occurrences we would expect for a particular code in a particular class is the probability associated with that code times the number of triplets associated with that class.

The correlation function is based on the number of triplets that a particular code occurs with in a particular class and the number of triplets we would expect it to occur within that class if it were randomly distributed. Intuitively, if a code occurs in a triplet the number of times we would expect, given the size of a particular semantic class, then the code is uncorrelated with respect to that class. If the code occurs in a triplet more than expected, it is positively correlated. The precise CORR function is actually the probability that a code would occur fewer times than actual, minus the probability that it would occur more times than actual:

$$\begin{aligned}&\mathrm{CORR}(\mathrm{code},\mathrm{class}) =\\ &\quad P(n < \mathrm{Count}(\mathrm{code},\mathrm{class})|\mathrm{code},\mathrm{class}) - P(n \geq \mathrm{Count}(\mathrm{code},\mathrm{class})|\mathrm{code},\mathrm{class})\end{aligned}$$

The values of CORR range over [-1.00,1.00]. A bimodal distribution is assumed in the calculation of $P(n|\mathrm{code},\mathrm{class})$.

As an example, consider the class 9.1 in (Levin, 1993) ("Put Verbs"). This class has 12 entries. 7 of these entries contain the 'X9' code. If 'X9' had been randomly distributed to this class, we would expect to find 1 occurrence of it. The CORR for 'X9' in the 9.1 class is 1.00, a very strong indicator that this code is applicable to this class. On the other hand, 'N' occurs on 7 entries in class 9.1. But we would expect it to occur 8 times if it were randomly distributed to this class. Consequently, the CORR for 'N' is low, -0.31; it is essentially uncorrelated with this class.

The method described above allows strong correlations to be posited even if only a small number of entries in a class contain a given code. For instance, the semantic class 9.7 ("Spray/Load" Verbs) has 44 entries, 11 of which occur with the LDOCE code 'T1-WITH'. Even though only one fourth of the entries have this code, the code has a CORR value of 1.00.

We used the CORR function as part of an experiment that relied on Bayes equation to predict a semantic class given a set of LDOCE codes:

$$M(\mathrm{class}|\mathrm{codes}) = \frac{P(\mathrm{codes}|\mathrm{class})P(\mathrm{class})}{P(\mathrm{codes})}$$

Given a set of codes, we can calculate $M(\mathrm{class}|\mathrm{codes})$ for every class and pick the class that has the highest probability. In order to do this we must be able to calculate the terms on the right side, $P(\mathrm{codes}|\mathrm{class})$, $P(\mathrm{class})$, and $P(\mathrm{codes})$. $P(\mathrm{class})$ is the number of entries in the class divided by the total number of entries. $P(\mathrm{codes})$ is the product of the independent probabilities of each code. $P(\mathrm{codes}|\mathrm{class})$ is the probability of a particular set of codes being produced from a member of the specified class. In order to calculate $P(\mathrm{codes}|\mathrm{class})$ we must be able to assign a value to $P(\mathrm{code}|\mathrm{class})$, the probability of a *single* given code appearing on a random entry in class. This probability depends on whether the code is correlated with the class, i.e., it is determined by the CORR function described earlier.

The CORR function allows us to introduce the notions of *marked* and *unmarked* classes. A class is *marked* if the CORR(code,class) is greater than some threshold (e.g., 0.8); it is *unmarked* otherwise. Once we have identified all the classes as marked or unmarked, we can determine the probability of a code appearing on an entry in a marked or unmarked class. If the class is marked for the code, then the probability of the code occurring is the number of times it occurred on entries in marked classes divided by the total number of entries in marked classes. If the class is unmarked, then the probability is the number of times it occurred on entries in unmarked classes divided by the total number of entries in unmarked classes.

To clarify this technique, we consider an example. There were 3828 total entries. The code 'D1' occurred 217 times in the entire set. Of those, 162 occurrences were on entries in classes that were marked for 'D1'. There were a total of 556 entries in the classes that were marked for 'D1'. Therefore, the probability of 'D1' occurring on a random entry is $\frac{217}{3828}$ (0.056688). The probability of it occurring on an entry in a class that is marked for 'D1' is $\frac{162}{556}$ (0.29136). The probability of 'D1' occurring on a class unmarked for 'D1' is $\frac{(217-162)}{(3828-556)}$ (0.016809). 'D1' is 17 times more likely to occur on an entry in a class that is marked for 'D1' that it is to occur elsewhere.

Using the above formulation of $M(\text{class}|\text{codes})$, we attempted to classify verbs. We calculated $M(\text{class}|\text{codes})$ for every class and recorded the 5 best scores. A subset of the raw results of this experiment is shown in Table 13. Correctly predicted classes are marked with a plus (+). Table 14 shows a comparison of the results of this experiment with the actual classification given in (Levin, 1993). Of the 3828 entries, 1807 (~47%) were correctly classified in one of the 5 best classes. Of the 2775 verbs, 1527 (55.0%) had at least one class correctly chosen.

The advantage of this approach over the brute-force technique is that it is robust; any new code pattern that is encountered will be classified on the basis of a correlation between syntactic information and Levin's semantic classes. It is possible that some combination of the brute-force technique and the robust approach will provide an optimal technique for lexical acquisition of verb-class information. We expect that more experimentation will be required in order to determine the threshold for application of one technique over another.

The final portion of this experiment will involve the incorporation of the enriched version of the thematic grids derived from Levin's semantic classes (e.g, see Table 9) into the Arabic lexicon. As before, our intent is to install these automatically (on the basis of English glosses) and to check the resulting entries by hand. This portion of the experiment will be conducted once the method of semantic classification has stabilized (i.e., after we have experimented with thresholds for a combination of brute-force and linguistically motivated techniques). Note, however, that one of our findings in the preliminary experiment (Section 2) was that human verification of reasonably accurate thematic grids is both more efficient and less error-prone than building thematic grids by hand. We expect those results to carry over to the second experiment.

Table 13. Sample Verb Classification Based on Robust Mapping from LDOCE Codes

Verb	**LDOCE Codes**	**Predicted Levin Classes**
abandon	T1 T1-TO N	22.4 29.7 42.1 22.3.c 13.2
acclaim	T1 T1-AS X1 N	33.a$^+$ 25.3 29.2 37.7 29.1
bike	I N	51.3.2 9.9 38 37.3 43.2
bleed	I I-FOR T1 T1-FOR	40.1.2$^+$ 41.1.1 31.3.c$^+$ 33.b 37.7
carpet	T1 N	31.1 51.3.2 9.9$^+$ 9.8$^+$ 22.4
clutter	T1 N	31.1 51.3.2 9.9 9.8$^+$ 22.4
conceal	T1 T4	33.b 31.2.b$^+$ 22.3.c 52
deem	X1 X1-TO-BE X7	37.7 29.3 9.5 37.6 29.2
depopulate	T1 WV5	31.1 9.8 5.3 45.4.f 47.8
enrich	T1 T1-BY T1-WITH	9.8$^+$ 47.8 1.1 23.1 33.b
free	T1 ADJ	47.8 31.1 2.2.a 13.5.2 45.4.c
funnel	L9 T1 X9 N	9.7 51.3.2 0.4.1 18.1 9.3$^+$
hang	I L9 T1 N	51.3.2 43.2 3.4 30.3 47.2
illuminate	T1 T1-WITH WV5	9.8 47.8 9.9 31.1 25.3$^+$
kick	I T1 X9 N	22.4 18.2 31.1 47.6 45.4.d
kidnap	T1	31.1 9.8 10.5$^+$ 10.6 33.b
ladle	T1 T1-INTO T1-OUT-OF N	10.6 9.7 21.2 31.1 9.3$^+$
lessen	I T1	31.1 45.4.d$^+$ 45.4.a 45.4.b 9.8
measure	I L1 T1 N	30.3 26.4 47.5.2 22.2.a 47.6
moan	I I-ABOUT T1 T5 N	38 37.3$^+$ 37.8 31.3.b 37.6
numb	T1 WV4 WV5 WV5-WITH ADJ	31.1$^+$ 45.4.a 9.9 10.5 51.3.2
panel	T1 T1-IN T1-WITH WV5 N	9.9$^+$ 9.8 47.8 22.2.a 25.1
pant	I I-AFTER I-FOR I3 T1 N	40.1.1$^+$ 35.6
peel	I T1 X9 N	22.4 18.2 31.1 47.6 45.4.d
pulverize	I T1	31.1 45.4.d 45.4.a 45.4.b 9.8
recreate	T1	31.1 9.8 10.5 10.6 33.b
shout	I T1 T5 N	38 37.3$^+$ 43.2 37.8 35.2
slumber	I N	51.3.2 9.9 38 37.3 43.2
swing	I I-FOR L7 L9 T1 X7 X9 N	51.3.1.a$^+$ 47.7 47.6$^+$ 50 26.5
tote	T1 N	31.1 51.3.2 9.9 9.8 22.4
unzip	T1	31.1 9.8 10.5 10.6 33.b
vomit	I T1 N	31.1 45.4.d 51.3.2 9.9 38
worsen	I T1	31.1 45.4.d$^+$ 45.4.a 45.4.b 9.8

Table 14. Semantic Classification Based on Linguistic Correlations

Choice					**Total** (out of 3828 entries)	**Total** (out of 2775 verbs)
1st	2nd	3rd	4th	5th		
814	379	262	196	156	1807 (47.2%)	1527 (55.0%)

4. Implications and Future Work

At first glance, the 55% success rate seems to be an insignificant result. Upon closer inspection, however, we see that this is not the case. A random classification algorithm would have a success rate of .5% for classifying a verb into one of 192 semantic classes. Thus, our algorithm is significantly better than a random classification algorithm. Moreover, our classification technique achieves this rate of success even in cases where the corresponding LDOCE pattern has not been encountered previously.

While the robust technique in Section 3.3 provides a reasonable foundation for fast lexical acquisition, we expect that it would be possible to improve the 55% success rate in future experiments by incorporating linguistic knowledge into the robust technique. In particular, we are currently investigating an approach that takes into account linguistic factors such as polysemy and positive and negative constraints on the syntax-semantics relation (Dorr and Jones, 1995). Polysemy is a problem in our current algorithm because we have assigned a single code pattern to each English verb, even in cases where a subset of the codes in a pattern corresponds to more than one verb meaning. We saw this in Section 3.3 for the case of *sleep* where two codes corresponding to two different verb meanings ('I' and 'T1') were collapsed together as a single code pattern. Multiple word meanings lead to a many-to-one mapping in both directions between target language items and their English glosses (i.e., many target senses for one English word or one target sense for many English glosses). If one sense maps to many senses in the target language, then the associated LDOCE codes will end up with the wrong semantic mapping and the word will be misclassified. Thus, it is crucial that the codes assigned in the initial language (English, in this case) be correctly separated into groups that correspond to independent word meanings. Currently, we are using post hoc checking techniques that provide the human checker with a small number of tests (based on syntactic diagnostics from Levin's book) to separate out incorrectly classified items. However, in future experiments, we will be separating LDOCE according to word meaning prior to activation of the acquisition program; we expect that this will significantly reduce the need for post hoc tests.

In addition to the addressing the problem of polysemy, we are currently investigating the benefits of applying linguistic knowledge concerning allowed and disallowed syntactic configurations for certain of the semantic classes. An analysis of the raw results of our previous experiments has revealed a possible correlation between LDOCE code pairs and combinations of syntactic behaviors, or "alternations" in the terminology of Levin. This correlation can be exploited in for more accurate prediction of membership or non-membership in a semantic class. An abbreviated list of possible correlations between Levin's alternations and LDOCE code pairs is given in Table 15.[20] The full set of correlations that we have discovered is presented in (Dorr and Jones, 1995).

The significance of this finding is that, with a comprehensive list, it is possible to impose not only positive constraints (which are the only type used in previous

Table 15. Proposed Correlations Between Levin's Alternations and LDOCE Code Pairs; Positive and Negative Constraints on Class Membership

Alternation	**Example**	**Code Pair**	**Pos/Neg**
Dative	Mary sent/said something to John Mary sent/*said John something	D1-TO/D1	37.4/*37.7
Conative	He nibbled/gobbled the cookie. He nibbled/*gobbled at the cookie.	T1/I-AT	39.2/*39.3
Causative/ Inchoative	She broke/designed the vase The vase broke/*designed	T1/I	45.1/*26.4
Causative/ Inchoative-Prep	Joe stood/lifted the books on(to) the table The books/*lifted stood on(to) the table	X9/L9	9.2/*9.4
Substance/ Source	Jelly oozed from sandwich The sandwich oozed jelly	T1/L9	43.4/—
Unspecified Object	John ate/dumped the apple John ate/*dumped	T1/I	39.1/*9.3
Zero-Related Nominal	a sparkle/*drown	N	43.1/*42.1

experiments) but also negative constraints. For example, the code pattern 'I T1' frequently occurs with verbs for which the Causative alternation is available, e.g., verbs in class 45.1 ("Break Verbs"), but not with verbs for which this alternation is barred, e.g., verbs in class 26.4 ("Create Verbs"). Given that Levin has a comprehensive listing of the semantic classes for which the Causative alternation is both allowed and disallowed, we can use the presence of the T1/I pair to rule out certain semantic classes that otherwise would have been considered by the acquisition techniques described in previous sections. In Table 15, positive and negative constraints on class membership are indicated by the absence and presence of the asterisk (*), respectively, in the column labeled **Pos/Neg**.[21] We expect that the use of both positive and negative constraints will raise the 55% success rate significantly.

A preliminary investigation indicates that such an approach will improve the results reported in Table 13. For example, the word "abandon" has a zero-related nominal (indicated by the LDOCE code 'N'), yet a semantic class that disallows zero-related nominals (42.1) is listed among the top five Levin-Based classes. Had we imposed the negative constraint shown in Table 15, i.e., that the code 'N' cannot be associated with a verb in class 42.1, we would have reduced the search space and thus raised the chances that the correct semantic class would appear in the top 5. Further experimentation will be required in order to determine the extent of improvement yielded by this enhanced approach.

An investigation into a possible correlation between LDOCE codes and the syntactic alternations of Levin was previously undertaken by Boguraev and Briscoe (1989), but only for the dative alternation. Their study involved an analysis of human judgments of meaning for those verbs containing dative-like codes (e.g., D1). The results of this study indicated that, on the basis of human judgments, membership in the semantic class of "Change of Possession" does not correlate well with the occurrence of dative-like codes. Note that this is a criticism of the Levin's assumptions about the semantic classification itself (which, in fact, might be more serious), not of the technique of mapping from LDOCE into the semantic classification. However, as the authors themselves indicate (p. 113), the dative alternation is only one, among many alternations that collectively make up the semantic classification of Levin. We take the view that the acquisition routine must necessarily examine *all* alternation possibilities (both positive and negative) before determining class membership, given that there is potentially a wide range of combinations and interactions. Moreover, Levin has not restricted the membership of dativizing verbs solely to the "Change of Possession" verb class; other classes are: "Having Verbs", "Send Verbs", "Slide Verbs", "Carry Verbs", "Drive Verbs", "Throw Verbs", "Transfer of a Message Verbs", etc.[22] We expect that a larger scale investigation of the code-class correlations based on more specialized code pairs (e.g., D1/D1-TO rather than just D1) across *all* alternations will produce more promising results, especially now that the semantic classification of Levin (1993) has been significantly refined since the time of this earlier study.

There are (at least) two difficulties beyond our control that will inevitably complicate any experiment that couples the LDOCE codes with the Levin classification. The first is simply that there are errors and omissions in the LDOCE. (This is also mentioned by Boguraev and Briscoe (1989).) For example, the verbs *grease* and *crown* are in the same verb class, yet only one of them, *crown*, contains the code 'WV5', which indicates that the '-ed' form can be used as an adjective. Clearly, the -ed form of the verb *grease* can also be used as an adjective, yet this information is not indicated in LDOCE. There is no single LDOCE code pattern corresponding to these two verbs; thus, an erroneous semantic classification is possible for at least one of them.

The second difficulty is that the syntactic tests that Levin uses to distinguish semantic classes are frequently much more subtle than the syntactic subcategorization information available in the LDOCE. In fact, we implicitly arrived at this conclusion already upon completion of the first experiment. That is, we determined that certain LDOCE codes (e.g., those corresponding to transitive and intransitive) simply are not sufficient for the generation of enriched (e.g., Levin-based) thematic grids. This implies that the LDOCE codes do not provide enough information to cover all of the combinations of syntactic behaviors required for distinguishing between the semantic classes in (Levin, 1993). For example, two crucial syntactic tests that Levin uses for distinguishing semantic classes are the Unspecified Object Alternation (e.g., that it is possible to say *John ate* as well as *John ate the apple*) and the Causative/Inchoative alternation (e.g., that it is possible to say *She broke the vase* as well as *The vase broke*). However, as can be seen in Table 15, these two tests are collapsed into a single LDOCE code pair T1/I. Thus, as far as the LDOCE is concerned, the verbs *eat* and *break* occur in the same semantic class. This type of collapsing is generally incorrect because, for example, the code 'I' might stand for an agent in once case, but a theme in another.

Regarding this last point, an area of future investigation that could prove useful is the analysis of dictionary definitions associated with verbs in the LDOCE. If we were to search the text of dictionary definitions for key words and phrases, it might be possible to pick out information that would help distinguish between two alternations, e.g., the two in Table 15 that are associated with the code pair T1/I.[23] Verbs in the "Change of State" class, which have the T1/I encoding, often have the word *cause* as part of their dictionary definition; the presence of *cause* indicates that T1/I should be correlated with the Causative/Inchoative alternation, not the Unspecified Object alternation. For example, the verb *break*, which has the T1/I encoding, is defined as follows: "to cause to separate into parts suddenly or violently." The verb *eat*, which also has the T1/I encoding, on the other hand, is defined as follows: "to take in through the mouth and swallow." Thus, it is possible to distinguish between the two possible correlations associated with the T1/I encoding. An attempt to pick out verbs in the "Change of State" class using a similar technique has been proposed by Fontenelle and Vanandroye (1989). A preliminary investigation indicates that the LDOCE would be useful in this regard as well.

5. Conclusions

We have argued that it is feasible to predict thematic-grid information from syntactic frames. The LDOCE has proven fruitful as a starting point for automatic incorporation of thematic grids into lexical entries. We have made use of a bilingual lexicon to incorporate the resulting thematic grids into an online Arabic dictionary. The resulting entries were shown to be highly accurate and the process of checking has proven to be more efficient than manual entry from scratch.

We have further experimented with methods that combined LDOCE codes with Levin's verb classes (Levin, 1993), demonstrating that it is possible to arrive at a more enriched semantic classification of verbs. We showed that there is an absolute limit (upper bound) of 72% accuracy for a brute-force, pattern-based approach to semantic classification and that it is possible to approach this upper bound with a more robust technique based on fine-tuned statistical correlations. We further suggest the possibility of raising this yield by taking into account linguistic factors such as polysemy and positive and negative constraints on the syntax-semantics relation. In any case, human intervention will always be necessary for construction of a semantic classification from LDOCE.

One of the benefits of using the Levin-based semantic classification is that, just as we are able to construct a set of thematic grids for each class (as shown in Table 9), we are also able to associate a lexical-semantic representation with each verb that is semantically classified. These representations serve as input to a composition process that derives the interlingua (based on earlier work by Dorr (1993)) for the translation process. We are currently augmenting the Arabic lexicon to include lexical-semantic representations with verbal entries. This will be achieved by associating lexical-semantic templates with semantic classes in (Levin, 1993) and installing the resulting forms in the corresponding lexical entries for Arabic.

The cross-linguistic applicability of the techniques described in this paper has been demonstrated in our current work on lexicon construction for Arabic, Spanish, and Korean. The lexicons resulting from this research are currently used in the PRINCITRAN prototype MT system (Dorr, 1993; Dorr et al., 1994; Lin et al., 1994).

Acknowledgments

We would like to express our gratitude to Doug Jones and Clare Voss for helpful commentary and suggestions and also to three anonymous reviewers who provided substantial guidance toward shaping this article into its final form. The research reported herein was supported, in part, by Army Research Office contract DAAL03-91-C-0034 through Battelle Corporation, NSF NYI IRI-9357731 and Logos Corporation, NSF CNRS INT-9314583, Advanced Research Projects Agency and ONR contract N00014-92-J-1929, Alfred P. Sloan Research Fellow Award BR3336, Army Research Institute contract MDA-903-92-R-0035 and Microelectronics and Design, Inc., and the University of Maryland General Research Board.

References

Alshawi, H. 1989. Analysing the Dictionary Definitions. In B. Boguraev and T. Briscoe, editor, *Computational Lexicography for Natural Language Processing.* Longman, London, pages 153–169.

Boguraev, B. and T. Briscoe. 1989. Utilising the LDOCE Grammar Codes. In B. Boguraev and T. Briscoe, editor, *Computational Lexicography for Natural Language Processing.* Longman, London, pages 85–116.

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

Dorr, B.J., J. Hendler, S. Blanksteen, and B. Migdalof. 1994. Use of LCS and Discourse for Intelligent Tutoring: On Beyond Syntax. In M. Holland and J. Kaplan and M. Sams, editor, *Intelligent Language Tutors: Balancing Theory and Technology.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Dorr, B.J. and D. Jones. 1995. Automatic Extraction of Semantic Classes from Syntactic Information in Online Resources. Technical Report UMIACS/CS TR, Institute for Advanced Computer Studies, University of Maryland, College Park, MD.

Dorr, B.J., D. Lin, J. Lee, and S. Suh. 1994. A Paradigm for Non-head-driven Parsing: Parameterized Message-Passing. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Farwell, D., L. Guthrie, and Y. Wilks. 1993. Automatically Creating Lexical Entries for ULTRA, a Multilingual MT System. *Machine Translation*, 8(3).

Fillmore, C.J. 1968. The Case for Case. In E. Bach and R.T. Harms, editor, *Universals in Linguistic Theory.* Holt, Rinehart, and Winston, pages 1–88.

Fontenelle, T. and J. Vanandroye. 1989. Retrieving Ergative Verbs from a Lexical Data Base. *Dictionaries*, 11:11–39.

Grimshaw, J. 1990. *Argument Structure.* MIT Press, Cambridge, MA.

Gruber, J.S. 1965. *Studies in Lexical Relations.* Ph.D. thesis, Information Science, Massachusetts Institute of Technology, Cambridge, MA.

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

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

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

Lin, D., B.J. Dorr, J. Lee, and S. Suh. 1994. A Parameter-Based Message-Passing Parser for MT of Korean and English. In *Proceedings of the Association for MT in the Americas Conference on Partnerships in Translation Technology, Columbia, MD*, pages 149–156, Columbia, MD.

Lonsdale, D., T. Mitamura, and E. Nyberg. 1995. Acquisition of Large Lexicons for Practical Knowledge-Based MT. *Machine Translation*, 9(3).

Montemagni, S. and L. Vanderwende. 1992. Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries. In *Proceedings of Fourteenth International Conference on Computational Linguistics*, pages 546–552, Nantes, France.

Pesetsky, D. 1982. Paths and Categories. Unpublished MIT Ph.D. dissertation.

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

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

Sanfilippo, A. and V. Poznanski. 1992. The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Resources. In *Proceedings of the Applied Natural Language Processing Conference*, pages 80–87, Trento, Italy.

Weinberg, A., J. Garman, J. Martin, and P. Merlo. 1994. Principle-Based Parser for Foreign Language Training in German and Arabic. In M. Holland and J. Kaplan and M. Sams, editor, *Intelligent Language Tutors: Balancing Theory and Technology.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Wilks, Y., D. Fass, C.M. Guo, J.E. McDonald, T. Plate, and B.M. Slator. 1989. A Tractable Machine Dictionary as a Resource for Computational Semantics. In B. Boguraev and T. Briscoe, editor, *Computational Lexicography for Natural Language Processing.* Longman, London, pages 85–116.

Wilks, Y., D. Fass, C.M. Guo, J.E. McDonald, T. Plate, and B.M. Slator. 1990. Providing Machine Tractable Dictionary Tools. *Machine Translation*, 5(2):99–154.

Notes

1. See also Dorr et al. (1994) and Weinberg et al. (1994) for a description of a related application, foreign language tutoring, that relies heavily on the automatically acquired lexical entries for Arabic.
2. A hint is given (p. 143) that future investigation is headed in the direction of automatic extraction of "case roles", but no details are given.
3. By "impoverished," we mean that these are minimal thematic relations based on simplistic subcategorization frames, e.g., transitive and intransitive. While subcategorization frames are clearly useful for syntactic processing, they do not describe the full range of thematic possibilities that would be necessary for translating between two languages. This was the motivation for moving toward more enriched thematic roles, as developed in our second experiment.
4. A similar viewpoint is expressed by Lonsdale, Mitamura, and Nyberg (1995) who have developed a large-scale knowledge using a combination of automatic techniques and human verification/refinement in order to define 60,000 words for translation of equipment manuals.
5. Alpnet is a company in Utah that develops translation aids. Their software originally ran on IBM's VM operating system, and we ported it to run on DOS and UNIX.
6. Arabic is a highly inflected language, with inflections occurring at the front and back of roots as well as root internally. The Alpnet system uses a sophisticated morphological analyzer to derive the underlying form for a given input word. Alpnet is fast, accurate and reasonably complete. Additionally, Arabic, like Hebrew, is typically written without vowels; Alpnet can accept and correctly handle both vowelled and unvowelled text.
7. In addition to extracting verbs, other principle parts of speech were collected, including nouns, prepositions, adjectives and adverbs. These non-verb words were needed to parse the multi-word glosses.
8. An empty set of parentheses () indicates that the thematic role must necessarily be contained in a prepositional phrase, but the prepositional head is left unspecified. Note that, in cases where the thematic role is not associated with parentheses, there is no prepositional head.
9. In English, verbs that are both transitive and intransitive can be either 'ag ag_th' or 'th ag_th'. An example of 'ag ag_th' is *eat*. In *John eats lunch*, the verb is 'ag_th', while in *John eats*, it is 'ag'. An example of 'th ag_th' is *freeze*; in *the water froze*, the verb is 'th', while in *John froze the water*, it is 'ag_th'. There is no reliable way to automatically distinguish these types of verbs. For our application this was not problematic, however, because Arabic does not have verbs of the 'th' type, choosing instead to use the passive for sentences like *the water froze*. Thus, we marked all glosses with an encoding of 'I' as 'ag'. Note that a more comprehensive theory of thematic roles *must* allow for alternative thematic grids corresponding to strict transitives and intransitives, especially if we are to consider languages other than Arabic. The second experiment addresses this concern by introducing an enriched set of thematic grids based on an extensive semantic classification.

10. A reviewer commented that *adorn* should be considered transitive in English since it takes the argument *oneself* as shown in the table. Note, however, that what we are trying to classify is a single Arabic root form (i.e., the root j_m_l_(V,intr) in the Alpnet lexicon) which, in this case, corresponds to the full phrase *adorn oneself*, not the single verb *adorn*. The Arabic verb corresponding to *adorn oneself* is, indeed, intransitive as our program correctly predicted.
11. The presence of the adverb removes many of the possible verb readings in English because of the adjacency requirement for verbs and objects. For instance:

    *John worked energetically Bill.

    But adjacency is not a problem in Arabic because these phrases represent a single verb. As such, the example given above might be rendered as:

    John worked Bill energetically.

    Consequently, many of the codes in the examples seem anomalous because of this adjacency difference.
12. Two exceptions were words suffixed with '-ize', which were considered transitive, and phrases of the form "be ...", which were considered intransitive. The rules from the "parsed matches" were used in these cases.
13. A grid was "attested" if it was deemed acceptable by the Arabic speaker based on Arabic usage examples analogous to the English sentences from the LDOCE.
14. We recently experimented with the construction of a Spanish verb lexicon from scratch. This effort was a six-month project for only 3500 verb entries, typed in by a native speaker and checked for accuracy by a second person. Given the labor and time involved in data entry and error correction for this much smaller project, we believe our time/error estimate for manual construction to be accurate (perhaps even optimistic).
15. The thematic grids shown here conform to the 4 conventions given in Section 2.2. In particular, the comma is used to indicate optionality of the subsequent roles and a parenthesized preposition is used to indicate a thematic role contained in a phrase headed by that preposition.
16. It is assumed that the reader has a familiarity with (Levin, 1993), which explains, in detail, what each of the semantic class numbers mean.
17. Levin's classes are labeled with numbers ranging from 9 to 57. However, the actual number of semantic classes is 192 (not 46) due to minor class subdivisions under each major class.
18. This coverage includes 1791 single-word glosses and 1507 multi-word glosses. Note that the single-word glosses match one of the 2775 unique Levin verbs exactly (e.g., *work*), whereas the multi-word glosses (e.g., *work energetically*) match some combination of a Levin verb (e.g., *work*) and some other constituent (e.g., *energetically*). Thus, the number 3298 (the full coverage of Arabic glosses) is greater than 2775 (the number of unique Levin verbs).
19. Multiple codes are ordered alphabetically except for the codes 'N' and 'ADJ' which always occur last. We imposed the alphabetical ordering to avoid spurious mismatches, e.g., between 'I T1 N' and 'T1 I N'.
20. Although the "Zero-Related Nominal" alternation is associated with a single code ('N' for noun), not a pair of codes, one could think of this code as being paired implicitly with any of the verbal codes (e.g., 'T1') since this is a cross-category alternation.
21. We have not enumerated all positive and negative constraints associated with each alternation; see (Dorr and Jones, 1995) for a more exhaustive treatment. We note, however, that negative constraints are not necessarily associated with every alternation. For example, the "Substance/Source" alternation is only associated with class 43.4, i.e., "Substance Emission Verbs"; in this case, it makes no sense to impose any negative constraints since this alternation is so narrowly applicable.
22. It is possible that this was not yet known at the time of the study by Boguraev and Briscoe (1989), given that (Levin, 1993) was published three years later.
23. We are indebted to an anonymous reviewer for pointing out this possibility to us.