Parameterization of the Interlingua in Machine Translation

Bonnie Dorr*
Institute for Advanced Computer Studies
A.V. Williams Building, Room 3157
University of Maryland
College Park, MD 20742
e-mail: bonnie@umiacs.umd.edu

Abstract

The task of designing an interlingual machine translation system is difficult, first because the designer must have a knowledge of the principles underlying cross-linguistic distinctions for the languages under consideration, and second because the designer must then be able to incorporate this knowledge effectively into the system. This paper provides a catalog of several types of distinctions among Spanish, English, and German, and describes a parametric approach that characterizes these distinctions, both at the syntactic level and at the lexical-semantic level. The approach described here is implemented in a system called UNITRAN, a machine translation system that translates English, Spanish, and German bidirectionally.

1 Introduction

What makes the task of designing an interlingual machine translation system difficult is the requirement that the translator process many types of *language-specific* phenomena while still maintaining *language-independent* information about the source and target languages. Given that these two types of knowledge (language-specific and language-independent) are required to fulfill the translation task, one approach to designing a machine translation system is to provide a common language-independent representation that acts as a *pivot* between the source and target languages, and to provide a parameterized mapping between this form and the input and output of each language. This is the approach taken in UNITRAN, a machine translation system that translates English, Spanish, and German bidirectionally. The pivot form that is used in this system is a lexical conceptual structure (henceforth, LCS) (see Jackendoff (1983, 1990), Hale & Laughren (1983), Hale & Keyser (1986a, 1986b), and Levin & Rappaport (1986)), which is a form that underlies the source- and target-language sentences.

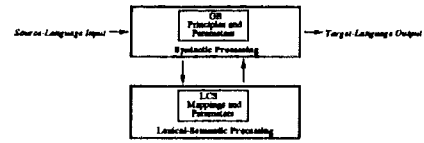


Figure 1: Overall Design of the UNITRAN System

The pivot approach to translation is called *interlingual* because it relies on an underlying form derived from universal *principles* that hold across all languages. Within this framework, distinctions between languages are accounted for by settings of *parameters* associated with the universal principles. For example, there is a universal principle that requires there to be a conceptual subject for each predicate of a sentence. Whether or not the conceptual subject is syntactically realized is determined by a parameter associated with this principle: the *null subject* parameter. This parameter is set to *yes* for Spanish (also, Italian, Hebrew, *etc.*) but *no* for English and German (also French, Warlpiri, *etc.*). The setting of the null subject parameter accounts for the possibility of a missing subject in Spanish and the incorrectness of a missing subject in English and German (except for the imperative form).

This paper argues that, not only should the syntactic component of a machine translation system be parameterized, but other components of a machine translation system would also benefit from the parameterization approach. In particular, the lexical-semantic component must be constructed in such a way as to allow principles of the lexicon to be parameterized. Thus, UNITRAN uses two levels of processing, syntactic and lexical-semantic, both of which operate on the basis of language-independent

*This paper describes research done at the University of Maryland Institute for Advanced Computer Studies and at the MIT Artificial Intelligence Laboratory. Useful guidance and commentary during the research and preparation of this document were provided by Bob Berwick, Gary Coen, Bruce Dawson, Klaudia Dussa-Zieger, Terry Gaasterland, Ken Hale, Mike Kashket, Jorge Lobo, Paola Merlo, James Pustejovsky, Jeff Siskind, Clare Voss, Amy Weinberg, and Patrick Winston.

knowledge that is parameterized to encode language-specific information (see figure 1).

Within the syntactic level, the language-independent and language-specific information are supplied, respectively, by the principles and parameters of *government-binding* theory (henceforth, GB) (see Chomsky (1981, 1982)). Within the lexical-semantic level, the language-independent and language-specific information are supplied by a set of general LCS mappings and the associated parameters for each language, respectively. The interface between the syntactic and semantic levels allows the source-language structure to be mapped systematically to the conceptual form, and it allows the target-language structure to be realized systematically from lexical items derived from the conceptual form. This work represents a shift away from complex, language-specific syntactic translation without entirely abandoning syntax. Furthermore, the work moves toward a model that employs a well-defined lexical conceptual representation without requiring a "deep" semantic conceptualization.

Consider the following example:

(1) (i) I stabbed John
(ii) Yo le di puñaladas a Juan
'I gave knife-wounds to John'

This example illustrates a type of distinction (henceforth called *divergence* as presented in Dorr (1990a)) that arises in machine translation: the source-language predicate, *stab*, is mapped to more than one target-language word, *dar puñaladas a*. This divergence type is *lexical* in that there is a word selection variation between the source language and the target language. Such divergences are accounted for by lexical-semantic parameterization, as we will see in section 3.

The following section of this paper will provide a catalog of syntactic divergences between the source and target languages. The set of parameters that are used to account for these divergences will be described. In the third section, we will examine the divergences that occur at the lexical-semantic level, and we will see how the parametric approach accounts for these divergences as well. Finally, we will turn to the evaluation and coverage of the system.

2 Toward a Catalog of Syntactic Divergences

Figure 2 shows a diagram of the UNITRAN syntactic processing component. The parser of this component provides a source-language syntactic structure to the lexical-semantic processor, and, after lexical-semantic processing is completed, the generator of this component provides a target-language syntactic structure. Both the parser and generator of this component have access to the syntactic principles of GB theory. These principles, which act as constraints (*i.e.*, filters) on the syntactic structures produced by the parser and the generator, operate on the basis of parameter settings that supply certain language-specific information; this is where syntactic divergences are factored out from the lexical-semantic representation.

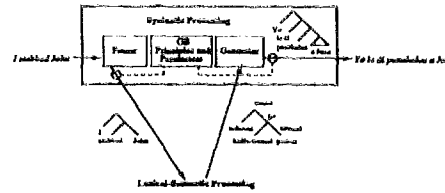


Figure 2: Design of the Syntactic Processing Component

The GB principles and parameters are organized into modules whose constraints are applied in the following order: (1) $\overline{\text{X}}$, (2) Bounding, (3) Case, (4) Trace, (5) Binding, and (6) θ. A detailed description of these modules is provided in Dorr (1987). We will look briefly at a number of these, focusing on how syntactic divergences are accounted for by this approach. Figure 3 summarizes the syntactic divergences that are revealed by the parametric variations presented here.[1]

2.1 Principles and Parameters of the $\overline{\text{X}}$ Module

The $\overline{\text{X}}$ constraint module of the syntactic component provides the phrase-structure representation of sentences. In particular, the fundamental principle of the $\overline{\text{X}}$ module is that each phrase of a sentence has a *maximal projection*, X-MAX, for a *head* of category X (see figure 4).[2] In addition to the head X, a phrasal projection potentially contains *satellites* α_1, α_2, β_1, β_2, γ_1, and γ_2, where α_1 and α_2 are any number of maximally adjoined adjuncts positioned according to the *adjunction* parameter, β_1 and β_2 are arguments (subjects and objects) ordered according to the *constituent order* parameter, and γ_1 and γ_2 are any number of minimally adjoined adjuncts positioned according to the *adjunction* parameter.[3]

[1] The syntactic divergences are enumerated with respect to the relevant parameters and modules of the syntactic component. The figure illustrates the effect of syntactic parameter settings on the constituent structure for each language. (In this figure, E stands for English, G for German, S for Spanish, and I for Icelandic.)

[2] The possibilities for the category X are: (V)erb, (N)oun, (A)djective, (P)reposition, (C)omplementizer, and (I)nflection. The Complementizer corresponds to relative pronouns such as *that* in *the man that I saw*. The Inflectional category corresponds to modals such as *would* in *I would eat cake*.

[3] This is a revised version of the $\overline{\text{X}}$-Theory presented in Chomsky (1981). The adjunction parameter will not be discussed here, but see Dorr (1987) for details.

Syntactic Divergence Examples		*Parameter*	*GB Module*
E, S: G:	V precedes object V follows object	constituent order	$\overline{X}$
E: S, G:	P stranding allowed No P stranding allowed	proper governors	Gov't
E, G: S:	Fronted question word beyond single sentence level not allowed Fronted question word beyond single sentence level allowed	bounding nodes	Bounding
E, G: S:	P not required before verbal object associated with clitic P required before verbal object associated with clitic	type of government	Case
E, G: S:	Subject required in matrix clause Subject not required in matrix clause	null subject	Trace
E, S, G: I:	Anaphor (*e.g., himself*) must have antecedent inside nearest dominating clause Anaphor (*e.g., sig*) may have antecedent outside nearest dominating clause	governing category	Binding
E: S: G:	No empty pleonastics allowed Empty pleonastics allowed Empty pleonastics in embedded clauses only	NDP	θ

Figure 3: Summary of Syntactic Divergences

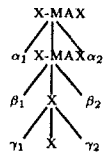


Figure 4: Phrase-Structure Representation

Given this general $\overline{X}$ phrase-structure representation, we can now "fit" this template onto the phrase structure of each language by providing the appropriate settings for the parameters of the $\overline{X}$ module. For example, the constituent order parameter characterizes the word order distinctions among English, Spanish and German. Unlike English and Spanish, German is assumed to be a subject-object-verb language that adheres to the verb-second requirement in matrix clauses (see Safir (1985)). Thus, for the sentence *I have seen him*, we have the following contrasting argument structures:

(2) (i) I have seen him
(ii) Yo he visto a él
'I have seen (to him)'
(iii) Ich habe ihn gesehen
'I have him seen'

The $\overline{X}$ module builds the phrase-structure from the general scheme of figure 4 and the parameter settings described above. The principles and parameters of the remaining modules are then applied as constraints to the phrase-structure representation. We will now examine each of the remaining modules in turn.

2.2 Principles and Parameters of the Government Module

Government Theory is a central notion to the Case and Trace modules. A familiar example of the government principle in English is that a verb governs its object.[4] We will examine the effect of this module in sections 2.4 and 2.5.

2.3 Principles and Parameters of the Bounding Module

The Bounding module is concerned with the distance between pairs of co-referring elements (*e.g.*, trace-antecedent pairs). The fundamental principle of the bounding module is that the distance between co-referring elements is not allowed to be more than one bounding node apart, where the choice of *bounding nodes* is allowed to vary across languages.

The bounding nodes parameter setting accounts for a syntactic divergence between Spanish and English (and German):

(3) (i) * Who$_i$ did you wonder whether t$_i$ went to school?[5]
(ii) ¿Quién$_i$ crees tú que t$_i$ fue a la escuela?

The reason (3)(i) is ruled out is that the word *who* has moved beyond two bounding nodes. It turns out that the corresponding Spanish sentence (3)(ii) is well-formed since the choice of bounding nodes is different and only one bounding node is crossed.

2.4 Principles and Parameters of the Case Module

The Case module is in charge of ensuring that all noun phrases are properly assigned abstract case (*e.g.*, nominative, objective, *etc.*). The Case Filter rules out any sentence that contains a non-case-marked noun phrase.

The notion of government is relevant to case assignment since an element assigns case only if it is a governing case-assigner. The setting of the *type of government* parameter for English, Spanish, and German characterizes the following divergences:

(4) (i) I saw Guille
* I saw to Guille
(ii) * Lo vi Guille
Lo vi a Guille[6]
(iii) Ich sah Guille
* Ich sah zu Guille

[4] See Dorr (1987) for a more formal definition of the government principle.

[5] If *who* is spoken emphatically, this sentence can almost be understood as an echo question corresponding to the statement *I wondered whether John went to school.*

2.5 Principles and Parameters of the Trace Module

After case has been assigned, the Trace module applies the *empty category* principle (ECP) which checks for proper government of empty elements. The ECP is parameterized by means of the *null subject* parameter. As discussed in section 1, the null subject parameter accounts for the null subject distinction between Spanish, on the one hand, and English and German on the other:

(5) (i) Yo vi el libro
Vi el libro
(ii) I saw the book
* Saw the book
(iii) Ich sah das Buch
* Sah das Buch

An additional parameter that is relevant to the Trace module is the *proper governors* parameter. The choice of proper governor accounts for preposition-stranding distinctions in the three languages:

(6) (i) [N-MAX What store]$_i$ did John go to t_i?[7]
(ii) * [N-MAX Cuál tienda]$_i$ fue Juan a t_i?
(iii) * [N-MAX Welchem Geschäft]$_i$ geht Johann zu t_i?

2.6 Principles and Parameters of the Binding Module

The Binding module is the final module applied before thematic roles are assigned. This module is concerned with the coreference relations among noun phrases, and it is dependent on the *governing category* parameter, which specifies that a governing category for a syntactic constituent is (roughly) the nearest dominating clause that has a subject. This parameter happens to have the same setting for English, Spanish, and German, but see Dorr (1987) for a description of other settings of this parameter (*e.g.*, for Icelandic) based on work by Wexler & Manzini (1986).

2.7 Principles and Parameters of the θ Module

The θ module provides the interface between the syntactic component and the lexical-semantic component. In particular, the assignment of *thematic roles* (henceforth θ-roles) after parsing leads into the construction of the interlingual form.

The fundamental principle of the θ module is the *θ-Criterion* which states that a lexical head must assign θ-roles in a unique one-to-one correspondence with the argument positions specified in the lexical entry for the head. One of the parameters associated with the θ module is the *nom-drop paradigm* (NDP) parameter (based on work by Safir (1985)). This parameter accounts for the distinction between English, on the one hand, and Spanish and German, on the other hand, with respect to the subject of an embedded clause:

(7) (i) * I know that was dancing
(ii) Yo sé que había un baile
'I know that (there) was a dance'
(iii) Ich weiß, daß getanzt wurde
'I know that (there) was dancing'

Once all θ-roles are assigned, the lexical-semantic component of the translator composes the interlingual representation for the source and target language. The next section will describe the lexical-semantic component, and it will show how this component accounts for a number of divergences outside of the realm of syntax.

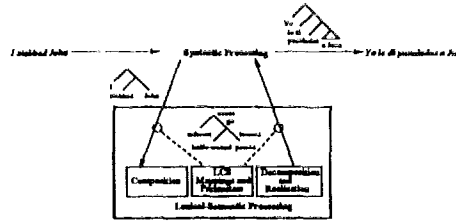


Figure 5: Design of the Lexical-Semantic Component

3 Toward a Catalog of Lexical-Semantic Divergences

Figure 5 shows a diagram of the UNITRAN lexical-semantic processing component. A detailed description of the *lexical conceptual structure* (LCS) which serves as the interlingua is not given here, but see Dorr (1990b) for further discussion.[8]

[6] As noted in Jaeggli (1981), animate objects (*e.g.*, *Guille*) are associated with a clitic pronoun (*e.g.*, *lo*) only in certain dialects such as that of the River Plate area of South America.

[7] The t_i constituent is a *trace* that corresponds to the noun phrase that has been moved to the front of the sentence.

[8] In general, the LCS representation follows the format proposed by Jackendoff (1983, 1990) which views semantic representation as a subset of conceptual structure. Jackendoff's approach includes such notions as Event and State, which are specialized into primitives such as GO, STAY, BE, GO-EXT, and ORIENT. As an example of how the primitive GO is used to represent sentence semantics, consider the following sentence:

(8) (i) The ball rolled toward Beth.
(ii) [Event GO ([Thing BALL],
[Path TO
([Position AT
([Thing BALL], [Thing BETH])])])]

This representation illustrates one dimension (*i.e.*, the *spatial* dimension) of Jackendoff's representation. Another dimension is the *causal* dimension, which includes

Lexical-Semantic Divergence Examples		Divergence Type (Parameter)	Associated Principle
E:	*enter*: John entered the house	Structural (*)	Linking rule
S:	*entrar*: Juan entró en la casa		
G:	*(hinein)treten*: Johann trat ins Haus hinein		
E:	*like*: I like Mary	Thematic (:INT, :EXT)	Linking rule
S:	*gustar*: Me gusta María		
G:	*gefallen*: Marie gefällt mir		
E:	*be*: I am hungry	Categorial (:CAT)	$\mathcal{CSR}$
S:	*tener*: Yo tengo hambre		
G:	*haben*: Ich habe Hunger		
E:	*like*: I like eating	Demotional (:DEMOTE)	Linking rule
S:	*gustar*: Me gusta comer		
G:	*gern*: Ich esse gern		
E:	*usually*: John usually goes home	Promotional (:PROMOTE)	Linking rule
S:	*soler*: Juan suele ir a casa		
G:	*gewöhnlich*: Johann geht gewöhnlich nach Hause		
E:	*stab*: I stabbed John	Conflational (:CONFLATED)	Linking rule
S:	*dar*: Yo le di puñaladas a Juan		
G:	*erstechen*: Ich erstach Johann		

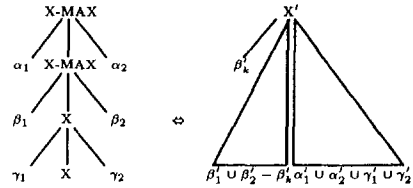
Figure 6: Summary of Lexical-Semantic Divergences

What is important to recognize about this processing component is that, just as the syntactic component relies on parameterization to account for source-to-target divergences, so does the lexical-semantic component. The parameterization of this component is specified by means of language-specific lexical override markers associated with the LCS mapping between the syntactic structure and the interlingua.

We will look briefly at the principles and parameters of the lexical-semantic component, focusing on how a number of divergences are accounted for by this approach. Figure 6 summarizes the lexical-semantic divergences that are revealed by the parametric variations presented here.[9]

the primitives CAUSE and LET. A third dimension is introduced through the notion of *field*. This dimension extends the semantic coverage of spatially oriented primitives to other domains such as Possessional, Temporal, Identificational, Circumstantial, and Existential.

[9]The divergences are enumerated with respect to the relevant principles and parameters of the lexical-semantic component. In contrast to the summary of syntactic divergences in figure 3, which enumerates the effect of syntactic parameter settings on constituent structure, the list of divergences presented here is specified in terms of the effect of LCS parameter settings on the realization of specific lexical items.



1. Syntactic specifier ($\beta_k \in \beta_1 \cup \beta_2$)
   $\Leftrightarrow$ Logical subject (β'_k)
2. Syntactic complements ($\beta_1 \cup \beta_2 - \beta_k$)
   $\Leftrightarrow$ Logical arguments ($\beta'_1 \cup \beta'_2 - \beta'_k$)
3. Syntactic adjuncts ($\alpha_1 \cup \alpha_2 \cup \gamma_1 \cup \gamma_2$)
   $\Leftrightarrow$ Logical Modifiers ($\alpha'_1 \cup \alpha'_2 \cup \gamma'_1 \cup \gamma'_2$)
4. Syntactic head (X)
   $\Leftrightarrow$ Logical head (X')

Figure 7: LCS Linking Rule Between the Syntactic Structure and the Interlingua

LCS Type	*Syntactic Category*
EVENT	V
STATE	V
THING	N
PROPERTY	A
PATH	P
POSITION	P
LOCATION	ADV
TIME	ADV
MANNER	ADV
INTENSIFIER	ADV
PURPOSE	ADV

Figure 8: $\mathcal{CSR}$ Correspondence Between LCS Type and Syntactic Category

3.1 Principles and Parameters of the Lexical-Semantic Component

The algorithm for mapping between the syntactic structure and the interlingua relies on the output of θ-role assignment (in the analysis direction) and feeds into θ-role assignment (in the synthesis direction). The θ-roles represent positions in the LCS representations of lexical entries associated with the input words. Thus, the construction of the interlingua is essentially a unification process that is guided by the pointers left behind by θ-role assignment.

The mapping, or *linking rule* between the syntactic positions and the positions of the LCS representation is shown in figure 7. In terms of θ-role assignment, the phrasal head X assigns θ-roles corresponding to positions in the LCS associated with X'. For example, the syntactic subject β_k is assigned the logical subject position β'_k in the LCS. Once all these roles have been assigned, the interlingual representation is composed simply by recursively filling the arguments of the predicate into their assigned LCS positions.

In addition to the LCS linking rule, there is another general rule associated with the lexical-semantic component: the *canonical syntactic representation* ($\mathcal{CSR}$) function. This function associates an LCS type (*e.g.*, THING) with a syntactic category (*e.g.*, N-MAX) (see figure 8).

The LCS Linking rule and the $\mathcal{CSR}$ function are the two fundamental principles of the lexical-semantic component. In order to account for lexical-semantic divergences, these principles must be parameterized. In general, translation divergences occur when there is an exception to one (or both) of these principles in one language, but not in the other. Thus, the lexical entries have been constructed to support parametric variation that accounts for such exceptions. The parameters are used in lexical entries as overrides for the LCS linking rule and $\mathcal{CSR}$ function. We will now examine examples of how each parameter is used.

3.1.1 '*' Parameter

The '*' parameter refers to an LCS position that is syntactically realizable in the surface sentence. This parameter accounts for *structural divergence*:

(9) (i) John entered the house
(ii) Juan entró en la casa
'John entered (into) the house'

Here, the Spanish sentence diverges structurally from the English sentence since the noun phrase (*the house*) is realized as a prepositional phrase (*en la casa*). In order to account for this divergence, the lexicon uses the * marker in the LCS representation associated with the lexical entries for *enter* and *entrar*. This marker specifies the phrasal level at which an argument will be projected: in the Spanish lexical entry, the marker is associated with an LCS position that is realized at a syntactically higher phrasal level than that of the English lexical entry.

3.1.2 :INT and :EXT Parameters

The :INT and :EXT parameters allow the LCS linking rule to be overridden by associating a logical subject with a syntactic complement and a logical argument with a syntactic subject. A possible effect of using these parameter settings is that there is a subject-object reversal during translation. Such a reversal is called a *thematic divergence*:

(10) (i) I like Mary
(ii) Me gusta María
'Mary pleases me'

Here, the subject of the source-language sentence, *I*, is translated into an object position, and the object of the source-language sentence *María* is translated into a subject position. In order to account for this divergence, the lexicon uses the :INT and :EXT markers in the LCS representation associated with the lexical entries for *gustar*. The English lexical entry does not contain these markers since the LCS linking rule does not need to be overridden in this case.

3.1.3 :CAT Parameter

The :CAT marker provides a syntactic category for an LCS argument. Recall that the $\mathcal{CSR}$ function maps an LCS type to a syntactic category (see figure 8). When this mapping is to be overridden by a lexical entry, the language-specific marker :CAT is used.

This parameter accounts for *categorial divergence*:

(11) (i) I am hungry
(ii) Ich habe Hunger
'I have hunger'

Here, not only are the predicates *be* and *haben* lexically distinct, but the arguments of these two predicates are categorially divergent: in English, the argument is an adjectival phrase, and, in German, the argument is a noun phrase. The :CAT marker is used in the German definition to force the PROPERTY argument to be realized as a noun rather than an adjective. Thus, the $\mathcal{CSR}$ function is overridden during realization of the word *Hunger* in this example.

3.1.4 :DEMOTE and :PROMOTE Parameters

The :DEMOTE and :PROMOTE markers, like the :INT and :EXT markers, allow the LCS linking rule to be overridden by associating a logical head with a syntactic adjunct or complement. These parameters account, respectively, for *demotional divergence*:

(12) (i) I like to eat
(ii) Ich esse gern
'I eat likingly'

and *promotional divergence*:

(13) (i) John usually goes home
(ii) Juan suele ir a casa
'John tends to go home'

In the first case, the English main verb *like* corresponds to the adjunct *gern* in German, and the embedded verb *eat* corresponds to the main verb *essen* in German. In the second case, the English adjunct *usually* corresponds to the main verb *soler* in Spanish. These "head switching" divergences are accommodated analogously: the :DEMOTE marker is used in the lexical entry for *gern* and the :PROMOTE marker is used in the lexical entry for *soler*.

3.1.5 :CONFLATED Parameter

The sixth LCS parameter is the :CONFLATED marker. This marker is used for indicating that a particular argument need not be realized in the surface representation. This parameter accounts for *conflational divergence* as in the sentence *I stabbed John* (see (1) from section 1). In this example, the

Class of Verb	*Lexical Primitives*
Position	STAY-TEMP, STAY-LOC, BE-TEMP, BE-LOC
Change of Position	GO-LOC, GO-TEMP
Directed Motion	GO-LOC, GO-POSS
Motion with Manner	GO-LOC
Exchange	CAUSE-EXCHANGE
Physical State	BE-IDENT, STAY-IDENT
Change of Physical State	GO-IDENT
Orientation	ORIENT-LOC
Existence	BE-EXIST, GO-EXIST, STAY-EXIST
Circumstance	BE-CIRC, GO-CIRC, STAY-CIRC
Range	GO-EXT-IDENT, GO-EXT-TEMP, GO-EXT-LOC
Intention	ORIENT-CIRC, ORIENT-TEMP
Ownership	BE-POSS, STAY-POSS
Psych State	BE-IDENT
Perception and Communication	HEAR-PERC, SEE-PERC
Mental Process	BE-PERC, GO-PERC
Cost	ORIENT-IDENT
Load/Spray	GO-LOC
Contact/Effect	GO-POSS

Figure 9: Coverage of Lexical-Semantic Primitives

argument that is incorporated in the English sentence is the KNIFE-WOUND argument since the verb *stab* does not realize this argument; by contrast, the Spanish construction *dar puñaladas a* explicitly realizes this argument as the word *puñaladas*. Thus, the :CONFLATED marker is associated with the KNIFE-WOUND argument in the case of *stab*, but not in the case of *dar*.

4 Evaluation and Coverage

One of the main criteria used for evaluation of the parameterization framework described here is the ease with which lexical entries may be automatically acquired from on-line resources. While testing the framework against this metric, a number of results have been obtained, including the discovery of a fundamental relationship between the lexical-semantic primitives and aspectual information. This relationship is crucial for demonstrating the success of the parameterization approach with respect to lexical acquisition. Details about the lexical acquisition model and results are presented in Dorr (1992).

We have already examined the syntactic and lexical-semantic coverage of the system (see figures 3 and 6 above). The linguistic coverage of the lexicon is summarized in figure 9.

5 Conclusion

The translation model described here is built on the basis of a parametric approach; thus, it is easy to change from one language to another (by setting syntactic and lexical switches for each language) without having to write a whole new processor for each language. This is an advance over other machine translation systems that require at least one language-specific processing module for each source-language/target-language pair.

The approach is interlingual: an underlying language-independent form of the source language is derived, and any of the three target languages, Spanish, English, or German, can be produced from this form. Perhaps the most important advance of UNITRAN is the mapping between the lexical-semantic level and the syntactic level. In particular, the system has been shown to select and realize the appropriate target-language words, despite the potential for syntactic and lexical divergences. The key to being able to provide a systematic mapping between languages is modularity: because the system has been partitioned into two different processing levels, there is a decoupling of the syntactic and lexical-semantic decisions that are made during the translation process. Thus, syntactic and LCS parameter settings may be specified for each language without hindering the processing that produces, and generates from, the interlingual form.

6 References

Chomsky, Noam A. (1981) *Lectures on Government and Binding*, Foris Publications, Dordrecht.

Chomsky, Noam A. (1982) "Some Concepts and Consequences of the Theory of Government and Binding," MIT Press.

Dorr, Bonnie J. (1987) "UNITRAN: A Principle-Based Approach to Machine Translation," AI Technical Report 1000, Master of Science thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Dorr, Bonnie J. (1990a) "Solving Thematic Divergences in Machine Translation," *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, University of Pittsburgh, Pittsburgh, PA, 127–134.

Dorr, Bonnie J. (1990b) "A Cross-Linguistic Approach to Machine Translation," *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Linguistics Research Center, The University of Texas, Austin, TX, 13–32.

Dorr, Bonnie J. (1992) "A Parameterized Approach to Integrating Aspect with Lexical-Semantics for Machine Translation," *Proceedings of the 30th Annual Conference of the Association for Computational Linguistics*, University of Delaware, Newark, DE.

Hale, Kenneth and M. Laughren (1983) "Warlpiri Lexicon Project: Warlpiri Dictionary Entries," Massachusetts Institute of Technology, Cambridge, MA, Warlpiri Lexicon Project.

Hale, Kenneth and Jay Keyser (1986a) "Some Transitivity Alternations in English," Center for Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA, Lexicon Project Working Papers #7.

Hale, Kenneth and J. Keyser (1986b) "A View from the Middle," Center for Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA, Lexicon Project Working Papers #10.

Jackendoff, Ray S. (1983) *Semantics and Cognition*, MIT Press, Cambridge, MA.

Jackendoff, Ray S. (1990) *Semantic Structures*, MIT Press, Cambridge, MA.

Jaeggli, Osvaldo Adolfo (1981) *Topics in Romance Syntax*, Foris Publications, Dordrecht, Holland/Cinnaminson, USA.

Levin, Beth and Malka Rappaport (1986) "The Formation of Adjectival Passives," *Linguistic Inquiry* 17, 623–662.

Safir, Ken (1985) "Missing Subjects in German," in *Studies in German Grammar*, Toman, Jindřich (ed.), Foris Publications, Dordrecht, Holland/Cinnaminson, USA, 193–229.

Wexler, Kenneth and M. Rita Manzini (1986) "Parameters and Learnability in Binding Theory," presented at the *Cognitive Science Seminar, MIT, September*, Cambridge, MA.

Zubizarreta, Maria Luisa (1987) *Levels of Representation in the Lexicon and in the Syntax*, Foris Publications, Dordrecht, Holland/Cinnaminson, USA.