The Case for a MT Developers' Tool with a Two-Component View of the Interlingua[1]

Bonnie Dorr and Clare R. Voss
Department of. Computer Science
University of Maryland
College Park, MD 20742
{bonnie,voss}@cs.umd.edu

Abstract

The interlingua (IL) in machine translation (MT) systems can be defined in terms of two components: (i) "lexical IL forms" within language-specific lexicons where each lexical entry has associated with it one or more lexical representations, and (ii) algorithms for creating and decomposing the instantiated "pivot" representation. Within this framework, we examine five different approaches to the level of representation for the lexical IL forms and then discuss a tool, ILustrate,[2] we are building to develop and evaluate different IL representations coupled with their corresponding translation algorithms.

1 Introduction

As discussed by Dorr and Voss (1993), MT theory has not yet addressed the issues surrounding how the interlingua (IL) in machine translation should be defined or evaluated. This paper provides a framework for addressing such issues. We take the IL in machine translation to be defined implicitly in two distinct ways: (i) declaratively and (ii) procedurally. We refer to the declarative portion of the IL as the "Lexical Component," consisting of the collection of IL entries from each natural language lexicon of an MT system. We refer to the procedural portion of the IL as the "Pivot-Form Component," consisting of the algorithms for creating and decomposing the full IL pivot form.

1.1 Interlocking of the IL Components

At IL definition time, the decisions concerning the two IL components are frequently interlocking, i.e., a change to one component drastically affects the functionality of the other, and vice versa. At MT run-time, by contrast, the relation between the components is fixed and one-way, with the pivot-form algorithms accessing the lexical IL forms. We briefly touch on an example of the interlocking problem here.

Consider the English words *about, by, down, in, off, on, out, over, under,* and up. These appear in two syntactic constructions (Lindner (1981)): the verb-particle construction (VPC, a complex with a verbal element and a particle) and the verb-prepositional phrase construction (VPP, a verb plus PP that may be an argument or an adjunct to the verb). These two constructions are illustrated here as (1) and (2), respectively:

[1] This research was supported, in part, by the Army Research Office under contract DAAL03-91-C-0034 through Batelle Corporation, by the National Science Foundation under grant IRI-9120788 and NYI IRI-9357731, and by the Army Research Institute under contract MDA-903-92-R-0035 through Microelectronics and Design,Inc.

[2] The acronym ILustrate stands for InterLingua Users' Support Tool, a Research And Testing Environment. We expect that the users of the tool will be MT researchers, MT system developers, and MT system evaluators.

(1) a. The intruder shot out the light.
b. The troops ran down the rations.
c. They looked up the address.

(2) a. The dog shot out the door.
b. The troops ran down the hill.
c. They looked up the chimney.

Note that a parser, using syntax alone, cannot determine the correct interpretation of these sentences. It could arbitrarily give just one analysis for the NP-V-P-NP sequences, or it could overgenerate and produce two or more analyses, including possibly one VPC and one VPP for each of the above sentences.

The interlocking problem arises when the developers of an interlingua start to build the lexical IL forms for the lexicon and then attempt to write pivot-form algorithms such that they are compatible with the lexical IL forms.[3] During algorithm development, they must take into account the range of sentential contexts, including the VPCs and VPPs, where spatial prepositions (as listed above) may appear; this often forces them to revise decisions about the IL forms placed in the lexicon.

Alternatively, an IL developer might start by focusing on the algorithms in the Pivot-Form Component. One option is to consider a strictly compositional approach in which case the developer might be forced to build IL forms for verbs first as the "seeds" of the pivot construction (or deconstruction) process; the IL forms for the particles might then be built later so that the final VPC form contains no overlapping substructures (as in a jigsaw puzzle). Another option would be to consider algorithms that permit operations beyond strict composition; in this case, the developer might be forced to build IL forms for the verbs and particles in such a way that they "link together" in an overlapping (possibly redundant) fashion.

In short, the choice of pivot-form algorithms affects the properties of the lexical IL forms and vice versa. Without IL prototyping support tools that allow for rapid reconstruction and testing of new lexical IL forms or pivot-form algorithms, the interlocking problem will remain in IL-based MT systems.

1.2 Formalization of the IL Components

Although interlinguas have been developed at many research sites, there are currently no tools to assist MT developers in establishing a formalism for describing IL forms. Ideally, an IL description tool would provide a syntax (e.g., a set of rewrite rules) and a semantics (e.g., an interpretation of primitive terms with respect to a knowledge model); these would be used to define, produce, or analyze the IL forms for the entries in each of the language-specific lexicons as well as for the testing and development of full IL pivot forms.

We view the formal description of (and support tools for) the Lexical Component to be complementary to, yet distinct from, that of the Pivot-Form Component. That is, we do not want one IL description language that simultaneously characterizes and generates the IL forms: instead we want two description languages in order to decouple the descriptions of constituent (i.e., *analysis)* trees needed for the Lexical Component from the descriptions of the construction (i.e., *derivation)* trees needed for the Pivot-Form Component.[4] In so doing, we adopt a reasoning that parallels the

[3] This is one typical sequence of events in MT research. Most linguistically-motivated representations are developed by linguists independent of the processing mechanisms that will operate on those representations. Thus, MT researchers are likely to adopt the linguistic representations first because these have been formally specified and then, in a subsequent step, they develop the algorithms that operate on these representations, rather than the other way around. In practical systems, by contrast, it is rarely feasible to decide on representations beforehand.

[4] For a formal description of analysis and derivation trees, see K. Vijay-Shanker, D. Weir, and A. Joshi, "Charac-

findings of researchers in the realm of syntax: constituent trees are not necessarily able to represent all the information in derivation trees (Joshi (1994)).

Note that it is not our intent *here* to build a new, complex interlingua or even to select one specific IL form over another. Rather we are. arguing for a mechanism that allows developers to build interlinguas in two components and overcome the interlocking problems that arise in scaling up their systems. The goal is to create an environment for assessing the adequacy of the two components (i.e., the proposed lexical forms and pivot-form algorithms) in covering productive combinations of data not seen before.

As the structural diversity and complexity of lexical IL forms continues to vary and as each new MT research site creates its own interlingua, it has become clear that the next step is to develop support tools for formalizing and testing IL components. It is expected that such tools will also prove useful in multi-system evaluations.

2 Levels of Representation

In this section we focus on the declarative portion of the IL, the Lexical Component in IL-based MT systems. We examine five current research approaches to the level of representation for the lexical IL forms: lexical-textual, lexical-ontological, lexical-semantic, lexical-syntactic, and tiered. What all these approaches have in common is that they are pushing the limits of two traditional assumptions implicit in IL research. The first is that the lexical IL forms that feed into the Pivot-Form Component exist at one predefined depth, or level of representation, beyond which further analysis does not occur. (Indeed in the classic transfer and IL "pyramid" diagram in Hutchins & Somers (1992, p. 107), one can even "see" this depth of analysis metaphor.) The second implicit assumption is that adequate translation is achieved only through exhaustive coverage at a single level of representation (Nirenburg et al. (1992)).

2.1 Lexical-Textual IL Forms

In the MT system Mikrokosmos, lexical entries are subdivided into three zones, corresponding to syntactic, semantic, and text meaning representation (TMR) information (Levin and Nirenburg (1994)). The TMR language is the formal basis for the interlingua in Mikrokosmos. It defines the acceptable lexical IL forms (or lexical-textual forms) and, via composition and decomposition of those forms, it also defines the full range of pivot forms that may appear during translation.

The unique characteristic of the TMR-based interlingua is that it is a collection of *microtheories* of meaning. These microtheories include meaning facets such as aspect, modality, evidentiality, speech acts, reference, speaker attitudes, stylistics factors, temporal relations as well as a "who did what to whom" component of meaning. These microtheories or components of meaning, when taken together, give the TMR-based IL its expressive strength. What we see here is that various sorts of knowledge can be defined, composed, and decomposed within the same formalism. What is less clear however is how these microtheories are defined at the lexical level. We can ask within this framework, for example, whether microtheories have Lexical Components of their own. Furthermore, if they do, or of those that do, we can also ask how the microtheory-based meaning of a particular lexical item contributes to the full IL pivot form (eg., via strict compositionality or not).

---

terizing Structural Descriptions Produced by Various Grammatical Formalisms" in *Proceedings of the 25th Annual Meeting of the ACL* (1987).

2.2 Lexical-Ontological IL Forms

Among AI researchers working on multilingual and MT systems, one of the most strikingly non-minimal approaches to lexical IL representations built in an ontological or conceptual framework is the current work of DiMarco, Hirst and Stede (1993). Their approach has been to develop two-part definitions by splitting lexical meaning along two levels of representation: a "conceptual" level for meaning components that are language-independent, i.e. interlingual, configurations of concepts, roles and associated fillers, as well as a "linguistic" level for meaning components that are language-specific structures and features tuned to capture fine connotational and denotational distinctions. The conceptual components are stored in a KL-ONE style taxonomic knowledge base (a KB with pointers into it from the lexical entries) and the linguistic components are stored in the relevant lexical entries.

Applying a two-component view of the IL to this research, we can see that the ontology is a back-door way of building a relational IL lexicon — the lexical IL forms are placed in well-defined relations to one another, not directly within the MT lexicon data structures themselves, but rather in the KB. This is a first step in isolating the Lexical Component of the IL. The single ontology containing all the lexical IL forms presents a framework to explore the space of lexical IL forms, a prerequisite to a formalization of the Lexical Component.

2.3 Lexical-Semantic IL Forms

The MT system UNITRAN developed by Dorr (1993) takes the theory of Jackendoff (1983, 1990) as the basis for the interlingua. Specifically, Dorr developed a modified, computational version of Jackendoff's "lexical-conceptual structures" (LCSs) as the formalism for lexical IL forms (her RLCSs) and pivot IL forms (her CLCSs). Although Jackendoff embedded his work in a psychological framework and has argued that his theory's semantic structures are conceptual structures (i.e equating semantic and conceptual levels of representation), it should be noted that these assumptions do not carry over to Dorr's work. UNITRAN assumes only that the LCS formalism provides a "syntax" or notation for encoding the lexical and sentential semantics, i.e. the interlingua. In other words, UNITRAN is theory-neutral with respect to the relation of semantics and conceptual structures.

Dorr's approach differs dramatically with the one discussed in section 2.2. Her work assumes that each individual lexical IL form is a (i) single, connected, annotated graph, that is a (ii) language-specific, (iii) semantic structure (iv) located in the lexicon — in marked contrast to DiMarco et al.'s two-part structures where one language-independent conceptual component is located in an ontology and the other language-specific linguistic component is stored in the lexicon. Several limitations to the Lexical Component in UNITRAN's interlingua as a lexical-semantic interlingua are a function of the gaps in Jackendoff's theory that Dorr relied on. For example, lexical entries for quantifiers, *not, and,* pronouns, and (in)definite articles — all central to research in logical semantics — were not covered directly by Jackendoff.

2.4 Lexical-Syntactic IL Forms

Recent work by Nomura et al. (1994) has focused on the development of an interlingua at a lexical-syntactic level of representation. Their work draws on the formal linguistic research of Hale and Keyser (1993), using Lexical Relational Structures (LRSs) as the basis for lexical IL forms.

One of the stated goals of this approach is to delimit the space of LRSs available for the Lexical Component of the IL. They criticize Dorr's use of LCS theory, for example, on the grounds that the space of LCSs is not constrained and so does not allow for a properly constrained mapping between sentential syntactic forms and full LCS-based IL pivot forms.

The way that Nomura et al. are expecting to show the value of their delimited LRS space is in terms of a well-defined and constrained mapping between the LRSs and sentential syntactic structure. Indeed within Hale & Keyser's work, the LRSs are also called "lexical syntactic structures" — making it clear that the broader shared research agenda is to push their current syntactic formalism down from the sentential level into the lexical level.

2.5 Tiered Lexical IL Forms

The goal of the tiered model (Dorr and Voss (1993), Dorr, Voss, Peterson, and Kiker (1994)) is to decouple the notions of an interlingua as a computational language and as a level of representation. Consequently the tiered form contains information derived from several levels of representation. For example, each constituent structure within a tiered form is *typed,* i.e. has a type value paired with it. The types are a small set of ontological categories from a *conceptual* level of representation. Also, for each predicator in a tiered form, there is an associated *semantic field,* such as a *locational, temporal,* or *possessional* field. These reflect the view that the patterns of lexical semantic structures in a variety of fields are shared because they are based on mappings from a few basic fields, such as location (eg. Anderson (1971), Talmy (1988)). Syntactic information is also encoded indirectly in the lexical forms: for each sub categorization frame that a verb may appear in, there is a distinct tiered lexical IL form. This mapping between frame and form indirectly preserves the information that is present in syntactic alternations through translation.

The tiered approach challenges the notion that all of the deepest, i.e. conceptual, knowledge is available in the interlingua. The underlying problem is partly theoretical and partly practical. Theoretically the difficulty arises from the assumption that IL constituents are equivalent to conceptual categories. This implies incorrectly that the meaning of a sentence is a rich knowledge structure. If this were indeed the case, then what would be the basis for bounding that structure?[5] If we, as developers, are forced to design our IL representations on the basis of unbounded structures, we would face the practical limitation that no representation, even if called conceptual, would capture the full meaning of an item in an MT lexicon.

3 ILustrate: an InterLingua Users' Support Tool

As stated at the beginning of this abstract, MT theory has not yet addressed the issues surrounding how the interlingua in machine translation should be defined or evaluated. The goal of the previous section was to show the variation *across* current IL-based MT research systems by examining, in a very limited way, lexical IL forms in the Lexical Components of a few of these systems. Some variation among interlinguas also continues at development time *within* each particular approach. For example, we have found, in the course of developing lexical-semantic forms in UNITRAN and tiered lexical IL forms in LEXITRAN, that our efforts to scale up the lexicon are hampered by

[5] Indeed, as pointed out by M. Kay et al.(1994), a *lexicalized* event is but one viewpoint of a real world event: the British action of *slotting a ticket in the machine when one gets on a bus or a train* is in French *invalidate the ticket,* and in German *validate the ticket.*

the lack of software to support (i) each cycle of specifying these IL forms with their associated pivot-form algorithms, and (ii) each cycle of testing the forms and algorithms.

In this section we introduce our work with ILustrate, a software tool to support development work during IL specification and testing cycles. We are currently in the beginning stages of implementation and have found that the same algorithms used for parsing, recognition, and generation in areas outside of our specific focus (i.e., IL development tools) are, in fact, applicable to the design of ILustrate.[6]

One of the goals of ILustrate is to assist in the scaling up of IL-based MT systems as the lexical IL forms or the algorithms for the IL pivot forms are revised to handle new data. ILustrate, in accord with our two-component view of the IL in MT systems, has two Specification Modules, one for the Lexical Component and one for the Pivot-Form Component. In addition to the reasons presented in the above, we add here another practical reason for keeping this division.

If we look at the MT research in progress even *within* one approach from the last section, we see that (i) there is variation among researchers in their interpretation of a particular linguistic or conceptual theory for building the lexical IL forms, and (ii) there are limitations with respect to what phenomena are handled. For example, recently, Verrière (1994) has developed lexical IL forms for French following a lexical-semantic approach much like that of Dorr (1993). What we see however is that, although the forms are similar, there remain differences that make importing Verrière's French forms into another MT system, however similar, a time-consuming task that will require expertise in the IL representation of each system as well as a knowledge of French. So one practical reason for designing ILustrate with a separate specification module for the Lexical Component is that it helps identify what types of declarative lexical IL variation exist between two MT systems. If a well-defined mapping can be established between the grammars of two systems' lexical forms, then a conversion algorithm can be built to adjust these forms before run-time. That is, by virtue of being able (i) to delimit the variation as lexical and (ii) to define a mapping between grammars of each system's lexical forms, we can scale up one MT system with lexical data from another system.

Since all theories imported for use as a basis for MT interlinguas are incomplete, invariably there will be another source of discrepancy between two research groups working even within the same approach: how each chooses to extend that theory to handle data outside the scope of the theory will differ. For example, researchers who focus primarily on translations at the predicate-argument level (such as those working with Jackendoff's LCSs or Hale & Keyser's LRSs) will eventually have to scale up their formalism to cover logical semantic words, including boolean-logical words *and, or, not,* causal-logical words *if, then, because,* and quantifiers. The IL that includes this class of lexical entries must capture both their *inherent* semantic sense as well as their scope (or domain of locality) in *derived* forms. In a two-component view of the IL, the properties of logical operators can be specified declaratively in the Lexical Component as well as procedurally in the composition algorithms of the Pivot Component.[7] The MT developer must specify how their logical terms will be interpreted. Thus another reason for designing ILustrate with a separate specification module for the Pivot Component is that eventually an MT system will need to be scaled up to include

---

[6] We are extending general natural language processing algorithms to tasks involving the interlingua. For example, our approach to decomposition of the IL pivot form involves parsing with a tree automaton and composition of the IL pivot form involves dynamic programming.

[7] A comparable choice in syntactic theories exists between a declarative structural encoding of scope (eg. tree-adjoining grammars) and a procedural encoding of scope via movement (eg. government-binding grammars).

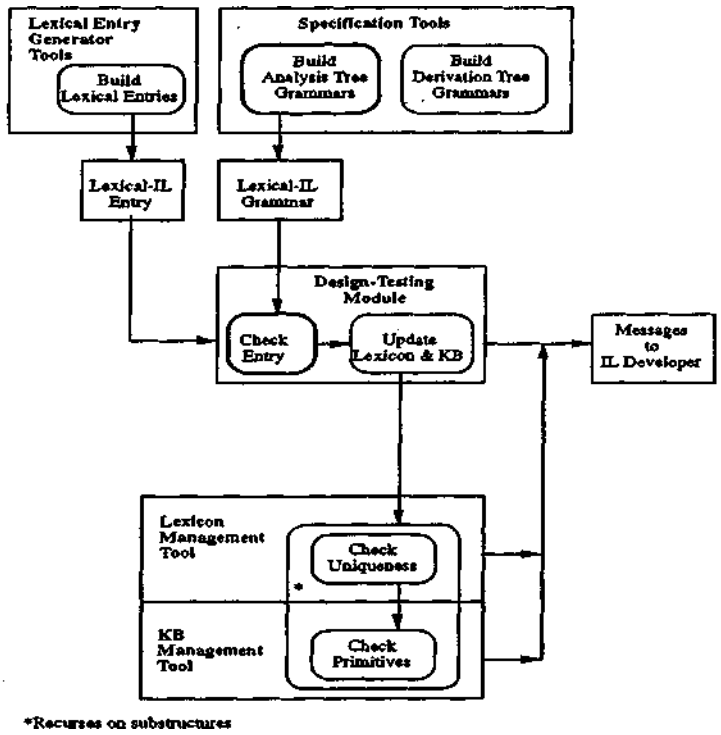

Figure 1: ILustrate Design: Lexical Component View

this class of entries and their scoping domains will need to be defined with respect to algorithms in that component.

In figure 1, the Specification Tools box (on the left side in the topmost row of boxes) contains two modules. The "build analysis tree grammars" tool is the module used by the MT developer to specify the grammar of their lexical component entries.[8] Once built, the Lexical-IL grammar can be used in the Design-Testing Module for a variety of functions. It may check the form of new lexical entries before they are added to the lexicon and knowledge base to ensure that their form is grammatically consistent with lexical entries already created.[9] The Lexical-IL grammar may also be used to read in entries of one form and generate a second set of entries that is consistent with a different grammar.[10] This, for example, is the application we need to work with Verrière's data (mentioned above).

The second module within the Specification Tools box (on the right side, the "build derivation tree grammars") is used by MT developers to specify the grammar of the Pivot Component operations in terms of derivation trees. Once built, the Pivot-IL grammar is brought into the Design-Testing Module of ILustrate during pivot form composition and for decomposition.[11] For

---

[8] The tree grammar formalism enables developers to specify a grammar for their IL as a set of trees, i.e. with no rewrite rules. Tree grammars are a more general formalism than the better-known classic grammars involving string rewrite rules.

[9] In some MT systems the lexicon and KB are combined, in others they are kept separate. The specification and testing modules for the Lexical Component of ILustrate are independent of this aspect of MT system design.

[10] Nothing in principle pre-empts adding a human checker into the loop to adjust the entries as well.
Given page limits, diagrams for the pivot form generation and decomposition could not be included.

example, the grammar guides the building of the pivot form so the developer can supply new lexical entries and then test and modify their interaction with the algorithms of the grammar. This speaks to the interlocking problem with spatial prepositions described in section 1.1: the developer can be prompted to supply new lexical entries (such as for phrases whose meaning is not strictly compositional) and then can test for the correct pivot forms using either the old or new lexical entries.

The Pivot-IL grammar, when brought into the Design-Testing Module during decomposition, enables the developer to submit a test IL form and have it disassemble the form into IL forms in order to check if these forms are available in the system lexicon. This provides, for example, a way of generating missing lexical-IL forms to be added to a new target language lexicon.

References

1. Anderson, J. (1971). *The Grammar of Case: Towards a Localist Theory,* Cambridge University Press, Cambridge, England.
2. DiMarco, C., G. Hirst and M. Stede (1993). "The Semantic and Stylistic Differentiation of Synonyms and Near-Synonyms," in Working Notes for the AAAI 1993 Spring Symposium on Building Lexicons for Machine Translation, Stanford University, CA, pp. 114-121.
3. Dorr, B. (1993). *Machine Translation: A View from the Lexicon,* MIT Press, Cambridge, MA.
4. Dorr, B. and C. Voss (1993) "Machine Translation of Spatial Expressions: Defining the Relation between an Interlingua and a Knowledge Representation System" in *Proceedings of AAAI,* Washington, DC.
5. Dorr, B., C. Voss, E. Peterson, and M. Kiker (1994). "Concept-Based Lexical Selection," in Working Notes for AAAI 1994 Fall Symposium on Knowledge Representation for Natural Language Processing in Implemented Systems, New Orleans, LA.
6. Hale, K. and J. Keyser (1993). "On Argument Structure and the Lexical Expression of Syntactic Relations," in K. Hale and J. Keyser (eds.), *The View From Building 20: Essays in Linguistics in Honor of Sylvain Bromberger,* MIT Press, Cambridge, MA.
7. Hutchins, W. J. and H. Somers (1992). *An Introduction to Machine Translation* Academic Press Inc., San Diego, CA.
8. Jackendoff, R.S. (1983). *Semantics and Cognition,* MIT Press, Cambridge, MA.
9. Jackendoff, R.S. (1990). *Semantic Structures,* MIT Press, Cambridge, MA.
10. Joshi, A. (1994). "From Strings to Trees to Strings to Trees...," Invited Talk, *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* Las Cruces, NM.
11. Kay, M., J. M. Gawron, and P. Norvig (1994). *Verbmobil: A Translation System for Face-to-Face Dialog* CSLI Lecture Notes Number 33, Stanford, CA.
12. Levin, L. and S. Nirenburg (1994). "The Correct Place Of Lexical Semantics in Interlingual MT," in *Proceedings of Fifteenth International Conference on Computational Linguistics* Kyoto, Japan.
13. Lindner, S. (1981). "A Lexico-Semantic Analysis of English Verb Particle Constructions with OUT and UP," University of California, San Diego.
14. Nirenburg, S. and J. Carbonell and M. Tomita and K. Goodman (1992). *Machine Translation: A Knowledge-Based Approach,* Morgan Kaufmann, San Mateo, CA.
15. Nomura, N., D. Jones, and R. C. Berwick (1994). "An Architecture for a Universal Lexicon: A Case Study on Shared Syntactic Information in Japanese, Hindi, Bengali, Greek, and English," in *Proceedings of Fifteenth International Conference on Computational Linguistics* Kyoto, Japan.
16. Talmy, L. (1985). "Lexicalization Patterns: Semantic Structure in Lexical Forms," in T. Shopen (ed.) *Grammatical Categories and the Lexicon,* University Press, Cambridge, England, pp 57-149.
17. Verrière, G. (1994). Manuel d'utilisation de la structure lexicale conceptuelle (LCS) pour représenter des phrases en français, Research Note, IRIT, Université Paul Sabatier, Toulouse, France.