Defining the Lexical Component in Interlinguas[1]

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

1 Introduction

As discussed by Dorr and Voss (1993, 1994), machine translation (MT) theory has not yet addressed the issues surrounding how the interlingua (IL) of a MT system should be defined or evaluated. This has a direct bearing on the decisions developers make with respect to the construction of a lexicon for MT. We view the IL as two distinct components: the declarative portion, which we call the "Lexical Component;" and the procedural portion, which we call the "Pivot-Form Component." The former component is a collection of entries from each natural language lexicon of the MT system. The latter component is the set of algorithms used to compose and decompose the full IL pivot form.[2]

The focus of this paper is on the definition of representations in the Lexical Component. We note, however, that the Lexical Component must be be tested in conjunction with the algorithms of the Pivot-Form Component. In particular, when IL representations are defined in the lexicon, 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. Interlocking problems may arise as the developers of an interlingua start to build the lexical IL forms for the lexicon and then attempt to write pivot-form algorithms that are compatible with the lexical IL forms. During algorithm development, they must then take into account the range of sentential contexts where each lexical item may appear; this often forces revisions to the IL forms in the lexicon.

We briefly examine the interlinguas in the lexicons of five current IL-based approaches to MT. Even with this limited review, the overall finding is clear: no consensus exists among MT researchers for defining (the equivalent of) the Lexical Component of an IL with respect to the levels of representation in an MT system. We provide a sketch of a tool, ILustrate,[3] currently under development for the design and evaluation of different lexical representations for interlingual MT.

2 Levels of Representation

This section provides an overview of the Lexical Component of five current IL-based MT systems. What all these approaches have in common is that they are pushing the limits

[1]This research was supported, in part, by the Army Research Office under contract DAAL03-91-C-0034 through Battelle Corporation, by the National Science Foundation under grants NYI IRI-9357731, NSF/CNRS INT-9314583, and NSF IRI-9120788, and by the Army Research Institute under contract MDA-903-92-R-0035 through Microelectronics and Design, Inc.

[2]We use the term "full IL pivot form" to refer to the complete IL form that is (i) created during the analysis phase of translation on the basis of the source language input text and (ii) deconstructed during the generation phase of translation.

[3]The acronym ILustrate stands for InterLingua Users' Support Tool, a Research And Testing Environment.

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 and 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. The microtheories, when taken together, give the TMR-based IL its expressive strength. We can ask within our framework, for example, whether microtheories have Lexical Components of their own.

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 is the current work of DiMarco, Hirst and Stede (1993), which focuses on the definition of meaning in terms of an ontology. In this work, lexical meaning is split between two levels of representation: a "conceptual" level for meaning components that are language-independent configurations of concepts, roles and associated fillers; and 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 and the linguistic components are stored in the relevant lexical entries.

The ontology developed in this work is a back-door way of building a relational IL lexicon — the lexical IL forms are placed in well-defined relations to one another in the KB. The single ontology containing all the lexical IL forms presents a framework to explore the space of lexical IL forms, a prerequisite for isolating and formalizing the Lexical Component of the IL.

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. Dorr developed a modified, computational version of Jackendoff's "lexical-conceptual structures" (LCSs) as the formalism for lexical IL forms and pivot IL forms. 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), Dorr assumes only that the LCS

formalism provides a "syntax" for encoding the lexical and sentential semantics, i.e., the interlingua.

In UNITRAN, 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 mentioned above. Several limitations to the Lexical Component in UNITRAN 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. This 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. This work presents a complementary view to that of Dorr's LCS theory in that it provides a constrained mapping between sentential syntactic forms and the LRS representation. Within Hale and Keyser's work, the LRSs are also called "lexical syntactic structures" — making it clear that the broader shared research agenda is to push the 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. Each constituent structure within a tiered form is *typed* by a small set of ontological categories. Each predicator has an associated *semantic field,* such as a *locational, temporal,* or *possessional* field. Syntactic information is also encoded indirectly in the lexical forms: for each subcategorization frame that a verb may appear in, there is a distinct tiered lexical IL form. This mapping between frame and form indirectly preserves subtle 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. This notion 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?[4] The practical limitation is that no representation captures the full meaning of an item in a MT lexicon.

3 ILustrate: A Support Tool for IL Lexicon Construction

In our own research, we have found that our efforts to scale up the lexicon are hampered by 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. Here we sketch out a few functions in ILustrate, a software tool to support development work during

[4]Indeed, as pointed out by 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.*

IL specification and testing cycles. 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. We are currently in the beginning stages of implementation (Dorr and Voss (1994)).

If we look at the MT research in progress even *within* one approach from the last section, we see (i) variation among researchers in their interpretation of a particular linguistic or conceptual theory for building the lexical IL forms, and (ii) limitations with respect to what phenomena are handled. For example, with respect to (i), recently Verrière (1994) has developed lexical IL forms for French following a lexical-semantic approach much like that of Dorr (1990). Although the forms are related, there remain differences that make importing Verrière's French IL forms into another MT system a time-consuming task requiring expertise in the IL representation and algorithms of each system as well as a knowledge of French. By having a separate specification module for the Lexical Component, ILustrate helps identify what types of declarative lexical IL variation exist between two MT systems. By being able to delimit the variation as lexical and define a mapping between grammars of each system's lexical forms, we can scale up one MT system with lexical data from another system.

With respect to (ii), invariably there will be another source of discrepancy between two research groups working even within the same approach: how they each choose 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 and 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 the full IL pivot forms. In ILustrate, our objective is to allow the IL developer the flexibility of respecifying their IL in either one or both of the Component specification modules.

In figure 1, the Specification Tools box at the top has two parts. The "build analysis tree grammars" tool is the module used by the MT developer to specify the grammar of their lexical component IL entries. Once built, the Lexical-IL grammar can be used in the Testing Module for a variety of functions. It may check the form of new lexical IL entries before they are added to the lexicon and knowledge base to ensure that their form is grammatically consistent with lexical entries already created.[5] 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.[6] This, for example, is the application we need to work with Verrièrre's data (mentioned above).

The second "build tree grammars" tool within the Specification Tools box (on the right side) is used by MT developers to specify the grammar of the Pivot-Form Component operations. Once built, the Pivot-IL grammar is brought into the Pivot-Form Component's Testing Module of ILustrate (not shown here) during pivot IL form composition and form decomposition. For example, this 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 in place. This speaks to the interlocking problem where the developer can be prompted to

[5] 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.

[6] Nothing in principle preempts adding a human checker into the loop to adjust the entries as well.

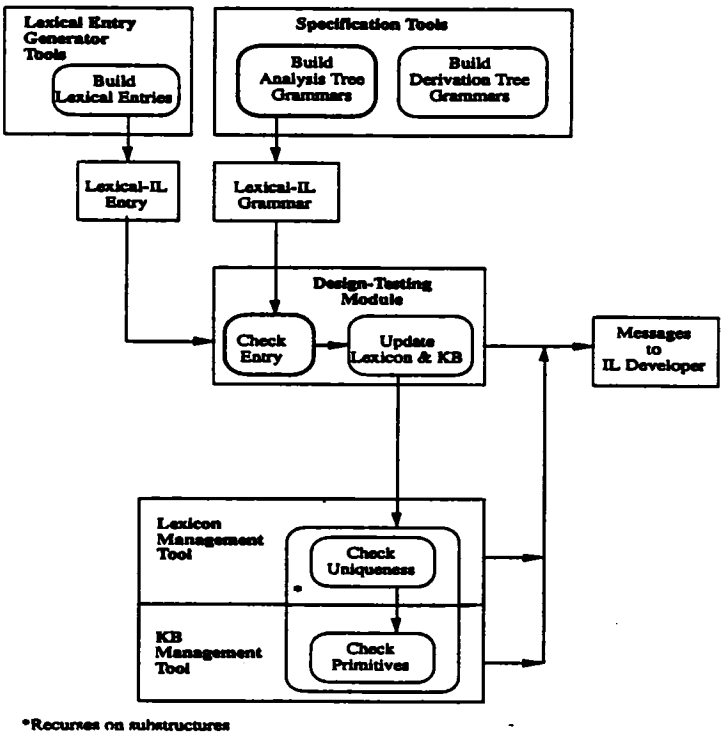

Figure 1: ILustrate Design: Lexical Component View

supply new lexical entries (such as for phrases whose meaning is non-compositional) and then can test for the correct pivot forms using either the compositional or non-compositional lexical entries. The Pivot-IL grammar, when brought into the 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.

4 Conclusion

In this paper we have focused both on the growing range of definitions for what counts as an interlingua in MT research and on the need for a software support tool that will help in future IL development work, particularly as it pertains to the lexicon. We view the current lack of consensus as a research opportunity to explore how the resources developed under the various approaches may be brought together to contribute to the overall progress in IL research. To that end, we are currently building a MT Development tool called ILustrate which aids in the development and evaluation of different IL representations for the lexicon.

References

1. 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.

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

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. and C. Voss (1994). "The Case for a MT Developers' Tool with a Two-Component View of the Interlingua" in *Proceedings of the First Conference of the Association for Machine Translation in the Americas* Columbia, MD.

6. 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.

7. 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.

8. Hutchins, W. J. and H. Somers (1992). *An Introduction to Machine Translation*, Academic Press Inc., San Diego, CA.

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

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

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. Nirenburg, S. and J. Carbonell and M. Tomita and K. Goodman (1992). *Machine Translation: A Knowledge-Based Approach*, Morgan Kaufmann, San Mateo, CA.

14. 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.

15. Schabes, Y. (1990). "Mathematical and Computational Aspects of Lexicalized Grammars," PhD Thesis, University of Pennsylvania, Philadelphia.

16. 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.