Machine Translation: A View from the Lexicon

Bonnie Jean Dorr
(University of Maryland)

Cambridge, MA: The MIT Press
(Artificial Intelligence Series, edited by
J. Michael Brady, Daniel G. Bobrow,
and Randall Davis), 1993, xx + 432 pp.
Hardbound, ISBN 0-262-04138-3, $45.00

Reviewed by
Daniel Radzinski
Tovna Translation Machines

1. Overview

Books describing novel approaches to machine translation (MT) are always welcome. This is all the more so when the approach is one not covered by general MT surveys such as those in Hutchins and Somers (1992) or Arnold et al. (1994). Bonnie Jean Dorr's *Machine Translation: A View from the Lexicon* is a book with a novel approach. It describes the interlingual MT system UNITRAN rooted in two Massachusetts-based frameworks of theoretical linguistics: Chomskyan principles-and-parameters government–binding (GB) theory for the syntactic component and Jackendovian lexical conceptual structure (LCS) for the lexical-semantic component, which also serves as the interlingua. The main claim is that cross-linguistic lexical-semantic divergences between source and target languages (at least across basic English, Spanish, and German) are of roughly only seven types, thus leading to a simple systematic translation mapping (relating the interlingua to the corresponding syntactic structures) parameterized by switches, with no language-specific rules.

Besides introductions, conclusions, and appendices, the book is organized into three parts encompassing UNITRAN's syntactic component, its lexical-semantic component, and application of the model. Chapter 1 is an introduction to the book. It briefly describes the basics of MT, including alternative approaches, and attempts to justify Dorr's parameterized interlingual principle-based design. It also begins a preliminary discussion of translation divergences, such as the lexical-semantic *categorial* type as in the English *I am hungry*, in which the predicate *hungry* is adjectival, compared with the German *Ich habe Hunger* ('I have hunger'), in which the corresponding *Hunger* is nominal.

Chapters 2 and 3 form the part dealing with the syntactic component. The former discusses the implementation of GB modules coupled with the parameters particular for each of English, Spanish, and German. The latter deals with the two-level morphological processor used in UNITRAN for analysis and generation.

Chapter 4, the first in the part dealing with the lexical-semantic component, is to a large extent a variant of Dorr (1993a). It describes the interlingual representation of UNITRAN. The chosen interlingua is an extended version of LCS, used also as a representation of lexical entries. Dorr justifies this choice as follows:

> [It] is suitable to the task of translating between divergent structures for two reasons: (1) it provides an *abstraction* of language-independent properties from structural idiosyncrasies; and (2) it is *compositional* in nature. (p. 95)

Also discussed is the mapping between the syntactic structure and the interlingua. In addition, an algorithm for LCS composition is introduced. Chapter 5 expands on the details of the systematic mapping between the interlingua and the syntax by specifying the relevant parameters and their values for English, Spanish, and German. A detailed translation example is given from the English *John broke into the room* to the Spanish *Juan forzó la entrada al cuarto* ('John forced entry to the room'), in which one sees how the constraints of the GB syntactic modules apply and how the mapping to and (de)composition of the interlingua take place. The example instantiates a solution to *lexical, structural,* and *conflational* divergences. Chapters 1, 2, and 5 together form a variant of Dorr (1993b). Chapter 6 describes the decomposition of the LCS interlingua and the syntactic generation of surface structure in the target language. Chapter 7 attempts to formalize and classify the different lexical-semantic divergence types and resolves these divergences by means of appropriate parameterizations.

Part III, "Application of the model," begins with Chapter 8, which presents the UNITRAN translation of a number of examples across the three languages, outlining some of the system's limitations. Chapter 9, the one I personally found the most interesting, describes current and future research on the application of UNITRAN. It proposes augmenting LCS with aspectual and temporal information, and it provides a model of lexical acquisition making use of such knowledge. Chapter 10 presents the conclusions of the book, and Appendices A–E offer various examples, rules, charts, screen dumps, and the like.

As mentioned previously, books of this type are always welcome, and I found it quite interesting in spite of its many technicalities (which are not necessarily exceedingly precise or explicit throughout). The extent to which the approach presented is promising is an altogether different question. In this regard, I have some doubts as to whether it is highly promising. I will elaborate this point in the following sections.

2. Linguistic Adequacy

It is unclear to me, at least from this book, whether UNITRAN meets the basic requirement of observational linguistic adequacy expected from any (serious) MT system. In other words, there are some apparently wrong linguistic/grammatical assumptions in UNITRAN. For example, the constraints from Case assignment form part of the syntactic GB Case module. Dorr claims (p. 222) that the English sentence *I gave to him the book* is unacceptable since the noun phrase *the book* does not get Case assigned by *gave* because *to him* intervenes between the two. Thus, if such a sentence were to be generated as a target output from, say, the Spanish source *Le di el libro,* it would be ruled out. It would likewise be ruled out as a source sentence, since it would fail the Case module constraints applied during analysis. The sentence is therefore judged unacceptable by the GB syntax used in the system.

But let us now consider the following sentence from the *Wall Street Journal*:[1]

> Mr. Richman's biggest victory so far was in helping to win passage of a 1984 California law that gives to deceased celebrities the same commercial rights enjoyed by the living. [8 Nov 1988]

According to Dorr and the GB framework she abides by, this sentence is to be judged

1 The *Wall Street Journal* material (copyright Dow Jones Inc.) here and below was extracted from the ACL/DCI CD-ROM I. All underlining is my own, in order to highlight the prepositional phrase between the verb and its direct object.

unacceptable for the same reason that *I gave to him the book* is judged unacceptable. However, it is perfectly fine, attestable, ordinary English. Along more general lines, Dorr claims that *John has eaten frequently breakfast* and *John has eaten in the morning breakfast* are unacceptable vis-à-vis *John has eaten breakfast frequently* and *John has eaten breakfast in the morning*, because "in English, an adverb or prepositional phrase may occur on the right side of a verb, but at the maximal level only (i.e., not between the verb and its object)" (p. 58). Yet the following, rather normal, *Wall Street Journal* examples show this to be blatantly wrong:

> Mr. Mitterrand said he hadn't asked the British leader to convey to the Russians French determination to keep its nuclear deterrent but he said she was in a position to speak for both of them on the issue. [25 Mar 1987]
>
> Mr. Littell's descriptions of the Russian Civil War of 1918–1920 convey sharply the carnage and brutality of that extremely bloody conflict. [13 June 1988]
>
> Mr. Creamer, a participant since 1984, likes the survey so much that his firm bought for clients 2,500 deluxe editions with gold-trimmed pages and an engraved cover. [30 Apr 1987]
>
> A few years ago, when an Amish friend needed cash to build a dairy barn, the Armstrongs bought from him a small, rocky patch of land on the crest of a wooded hill and built their home on it. [04 Dec 1987]
>
> Zapata Corp. said its bank lenders extended through April 30 the deferrals of payments and covenant waivers that were to expire last Saturday. [03 Mar 1987]
>
> BP extended by one day its $7.9 billion tender offer for Standard shares that it didn't previously own. [05 May 1987]
>
> His father instilled in him a commitment to public service, frugality and a love of fishing. [26 Oct 1987]
>
> He said Mr. McFarlane relayed to him a directive from President Reagan to keep the Contras together 'body and soul' after Congress suspended official aid to them in October 1984. [07 July 1987]
>
> In an opinion by then Chief Justice Warren Burger, the high court discussed at length the historical purpose of recognizing charitable organizations, ... [08 June 1987]

Furthermore, Dorr herself writes in the acknowledgments

> This book carries with it the memory of four family members, all of whom left us in one difficult year, ... (p. xx)

It is true that these sentences exhibit objects which are "heavier" than the adverbs or prepositional phrases (PPs), suggesting that some pragmatic, i.e., non-syntactic, heavy-constituent shifting to the end of the sentence is taking place here, thereby alleviating the processing load and eliminating potential PP attachment ambiguities that might arise. However, the following examples do not exhibit any substantially greater heaviness of the object when compared with the PP (though we would encounter some problems with anaphora interpretation if the object preceded the PP in the first example):

> Also recently, Dallas-based National Southwest Capital Group Inc. said it bought from Mr. Waldron and his family their 4.7% stake in Ocilla. [24 July 1987]
>
> Wilmington Trust Co. said it extended until 5 p.m. next Wednesday its offer to merge with Delaware Trust Co. [07 May 1987]

Some Eastern pilots plan to use the meetings to convey to the national pilot-union leadership their continued commitment to the strike. [07 Aug 1989]

Dorr may claim that "if we were to pick out a random sentence from a novel, or even a newspaper, it is highly likely that we would run into stylistic idiosyncrasies that would be too difficult to handle" (p. 307). We have indeed picked out newspaper sentences here, but these hardly exhibit any stylistic idiosyncrasies with respect to PP positioning. All that they, and Dorr's own sentence, seem to imply is that the grammatical premise ruling out *I gave to him the book, John has eaten frequently breakfast,* and *John has eaten in the morning breakfast* is simply wrong. These three sentences are syntactically fine, though communicationally odd. If one were to claim that all verb–PP–object sentences are ungrammatical and yet some are found acceptable owing to post-syntactic movement (i.e., at a purely syntactic level the acceptable ones are really verb–object–PP), then a serious MT system based on such a thesis would have to show an algorithm for getting from the abstract syntactic level to the concrete surface form. I would tend to consider an MT system translating sentences across abstract syntactic levels rather than across ordinary, human, natural languages as one outside the realm of empirical scientific research, and it is unclear to me whether we could actually refer to it as an MT system at all.

Another case related to the issue of observational linguistic adequacy has to do with the translation into Spanish of the German *Ich habe Hunger* ('I am hungry'). In UNITRAN, this translation ends up as either *Yo tenga hambre* or *Yo tengo hambre.* The form *tenga* is in the subjunctive mood, and hence *Yo tenga hambre* as a full sentence is unacceptable, whereas *tengo* is in the indicative mood, and hence *Yo tengo hambre* is acceptable. Dorr justifies the subjunctive *tenga* output "since the [German] verb *habe* can be interpreted as either subjunctive or indicative" (p. 295). But the entire sentence *Ich habe Hunger* can be interpreted only in the indicative! (Or so at least it seems to me.) Simply because the form *habe* at a word level can be also interpreted as subjunctive does not automatically mean that subjunctiveness must ultimately be transferred to Spanish. The subjunctive interpretation should actually be ruled out early in the analysis of the source, as the string is a full sentence after all. There is no excuse then for allowing here an unacceptable subjunctive *tenga,* even as a second possibility concomitant with the correct one. A serious MT system should not allow syntactically unacceptable target overgeneration—and mood is clearly a syntactic matter—simply because of morphological underspecification in the source.

3. Completeness

Besides observational linguistic adequacy, there is also the question of incompleteness. Every MT system is incomplete. The question is when is such a system so incomplete as to be of little significance for the overall MT endeavor? UNITRAN has a number of incompletenesses that are understandable. Here are a few examples:

- There are only 305 English verbal roots in its morphological lexicon (p. 84), when a reasonable basic system ought probably to begin with ten times that amount.
- Almost all of the examples tested in the system, or at least most of those presented in the book and in Dorr (1993a, 1993b), are sentences of less than ten words, when in fact average sentences in ordinary written language generally consist of at least three times that many words.

- Existential *there* sentences, e.g., *There is a man in the room*, are not currently handled (p. 220, fn. 20). It is not necessary to stress that such sentences form an extremely common, important, and central construction in the language.

These incompletenesses might suggest that we are dealing here with something of a "toyish" system. However, one must start somewhere, so it would make sense for the system to have such gaps in its early stages of development. This is particularly true if UNITRAN is meant to deal primarily with cross-linguistic divergences and competence-based problems and not necessarily with real true-to-life examples.

Another incompleteness has to do with context, pragmatics, and knowledge of the world. The system is in essence a syntactic and lexical-semantic one and so one does not expect it to deal much with these other difficult areas. Dorr, however, exhibits some ambivalence with respect to this issue. On the one hand, she indicates that "context is not part of the model" (p. 212, fn. 11). On the other hand, she does not allow *Ich fresse gern* to be generated in German as a possible translation of *I like to eat*, since the German verb *fressen* "requires an agent that is [a non-human] animal" (p. 208, fn. 9). Now, if the only difference between *essen* and *fressen* is that the latter requires its agent subject to be a non-human animal while the former lacks such a requirement, then we are indeed dealing here with a pure syntactic/lexical-semantic case. However, *fressen* in fact may have a human agent with a meaning akin to *eat like a pig* as in *Ich fresse nicht wie ein Schwein* ('I don't eat like a pig'). So the requirement is actually a pragmatic one, and hence UNITRAN should not really reject *Ich fresse gern* if context is indeed outside of its domain. A less incomplete system might give preference to *Ich esse gern* over *Ich fresse gern*, rather than reject the latter outright, by appealing, as Dorr herself suggests (pp. 159–162), to knowledge-based techniques. Another approach, as used for example in the system currently being developed by Tovna Translation Machines, would obtain the preference by appealing to various sorts of statistics, rather than to any particular knowledge-based representations.[2]

An additional incompleteness concerns metaphorical language. On this matter Dorr aptly indicates that "the problem of metaphor is well outside of the bounds assumed in the design of UNITRAN (and, for that matter, *most* machine translation systems)" (p. 309). This is fully understandable. She also claims that "metaphorical items must be mapped to the conceptual structures underlying their 'true' meanings before translation to the target language can proceed 'normally' " (p. 308). Consequently, UNITRAN translates *kill a process* into the odd Spanish *matar un proceso* ('[literally] kill a process'), rather than the more appropriate *acabar con un proceso* ('terminate a process'). The metaphorical use of *kill* in the sense of *terminate* does not hold for the Spanish *matar*. But if, as Dorr claims, we would need to map *kill* to the LCS of *terminate*, as this is its "true" meaning, before any translation can proceed "normally," then perhaps, at least at this stage of research and development, LCS is the wrong representation to use as an MT interlingua, given the vast use of metaphor in translatable natural language. A somewhat more direct mapping from *kill* to *acabar con* in this context would seem far more suitable than the LCS compositions and decompositions used in UNITRAN.

2 Such statistics could also serve as a "last resort" by giving very low priority to examples of overgeneration due to lack of collocational tests in UNITRAN. Thus the overgenerated examples in Spanish of **Juan fue en la casa* and **Juan entró a la casa* (p. 169, fn. 3) from *John entered the house* would end up having a very low statistical priority and would thus be ruled out if certain statistically based thresholds are set, with no need to appeal to any collocational tests.

4. Divergences

A third problem has to do with what I refer to as arbitrary choices. A number of problems may arise as the result of a particular choice taken. For example, Dorr discusses at length *demotional* and *promotional* divergences (pp. 176–183 and 268*ff*). An example of a demotional divergence is the English *I like to eat*, in which *like* is a main verb, compared with the German *Ich esse gern* ('I eat likingly'), in which *gern* is an adjunct adverb carrying with it the notion of 'liking.' An example of a promotional divergence is the English *John usually goes home*, in which *usually* is an adjunct adverb, compared with the Spanish *Juan suele ir a casa* ('John tends to go home'), in which *suele* is an inflected main verb. Both *usually* and *soler* carry with them the notion of 'habit.' Dorr argues for a distinction between these two kinds of divergences.[3] These divergences are prima facie problematic, as it is no easy task for a machine to translate a main verb in the source to some adjunct item in the target, or vice versa. Dorr devotes much discussion to the resolution of such divergences. But note that such divergences exist only if one considers *like* and *suele*, in the above examples, to be main verbs. Nothing bars *eat* or *ir* from serving as the main verbs, with *like* and *suele* respectively serving as their modal-like auxiliaries. In such a case, there really is no serious divergence, since the main verbs in the target and source are translations of each other. We would still have to deal with a categorial divergence between adjunct adverbs and auxiliary verbs, but this is far easier to handle. (In fact, the system currently being developed at Tovna makes use of such a representation and thus circumvents demotional or promotional divergences.) UNITRAN may be able to resolve a problem created because of a particular syntactic representation it has chosen to use, but the problem would never have arisen with a different representation.[4] Is then the GB version used by UNITRAN, or perhaps *any* GB version, an appropriate and promising syntactic framework for MT?[5]

5. Tense and Aspect

What does seem quite promising, at least prima facie, is the research reported in Chapter 9 on augmenting LCS, and hence UNITRAN, with information on aspect and tense. Here Dorr has been able to enrich LCS, in a parameterized way, with information exceedingly important for MT purposes. For example, a telicity parameter accounts for the subtle distinctions among *ransack*, *destroy*, and *obliterate* (p. 339*ff*). Furthermore, an empirical analysis of corpora has established a hitherto unclear crucial link between LCS representation and a well-known aspectual classification scheme (p. 341*ff*). This in itself is a nice result.

3 The distinction is not biased toward English, but rather is based on reasonably objective criteria.

4 Similarly, some of the divergences are due to a particular choice of translation. Thus if one chooses to translate the French *Il est probable que Jean viendra* into the English *Jean will probably come* (p. 236) rather than a more isomorphic *It is likely that Jean will come*, then one obtains, according to Dorr's classification, a promotional divergence. If, however, one matches the source sentence with the perfectly fine target *It is likely (that) Jean will come*, then we have no divergence at all. This point is also true if we match an English source, *Jean will probably come*, with a French target, *Jean viendra probablement*.

5 Some support for a negative answer to this question, as well as to the question concerning the appropriateness of LCS for MT, may come from Dorr's report to the effect that it took close to 3 hours to translate the German *Johann brach ins Zimmer ein* ('John broke into the room') into the Spanish *Juan forzó la entrada el cuarto* (p. 300 and p. 384). Three hours on *any* work station in order to translate a five-word sentence strongly suggests that the system is barking up the wrong tree somewhere.

6. Conclusion

Overall, the book will interest, in one way or another, anyone engaged in MT research and development. It will also serve linguists interested in seeing how GB and LCS can be put to use computationally. However, it remains to be seen whether the approach adopted in the book will ultimately lead to any significant progress in MT.[6]

References

Arnold, Douglas; Balkan, Lorna; Humphreys, R. Lee; Meijer, Siety; and Sadler, Louisa (1994). *Machine Translation: An Introductory Guide*. Oxford: NCC Blackwell.

Dorr, Bonnie J. (1993a). "The use of lexical semantics in interlingual machine translation." *Machine Translation* **7**(3):135–193.

Dorr, Bonnie J. (1993b). "Interlingual machine translation: A parameterized approach." *Artificial Intelligence* **63**:429–492.

Hutchins, W. John and Somers, Harold L. (1992). *An Introduction to Machine Translation*. London: Academic Press.

Daniel Radzinski received his doctorate in mathematical linguistics from Harvard University in 1990. In 1990–91, he was Visiting Assistant Professor for Computational Linguistics in the Department of Cognitive and Linguistic Sciences, Brown University. Since late 1991, he has been Senior Computational Linguist at Tovna Translation Machines, engaged in software and grammar development for a robust machine translation system that currently covers English, French, and Russian. Radzinski's address is Tovna Translation Machines Ltd., 28 Beit Ha'arava St., Jerusalem 93389, Israel; E-mail: dr@tovna.co.il

6 From an editorial standpoint, the book is quite good. I detected only eight editorial infelicities. Almost all of them are rather innocuous and easily correctable by the reader. The only blatant error is glossing the Spanish *Juan entro en la casa* as 'I saw to John' rather than 'John entered in(to) the house' (Figure p. 20, repeated p. 165).