Lojban as a Machine Translation Interlanguage in the Pacific

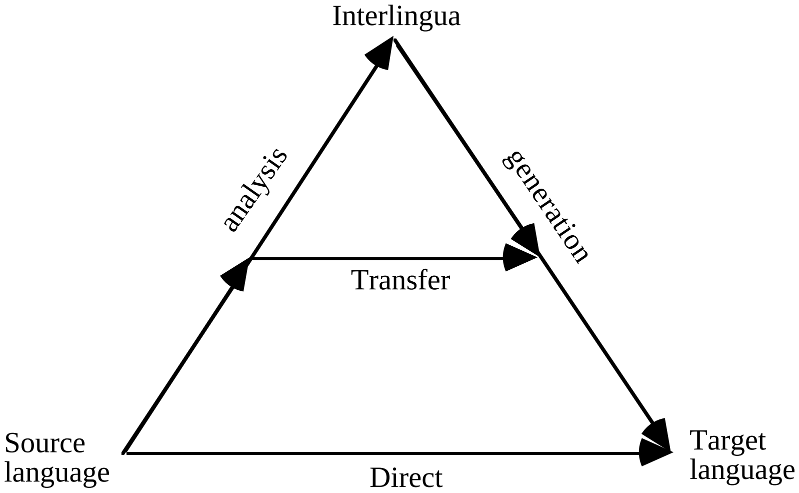
Nick Nicholas
Department of Linguistics & Applied Linguistics,
University of Melbourne
N.Nicholas@linguistics.unimelb.edu.au

1. Introduction

The dominant strategy currently employed in machine translation (MT) is *transfer:*[1] Under this strategy, the MT system derives an abstract representation of the source language text; this representation is then transferred into a corresponding abstract representation formed in terms of the target language, which is subsequently realised in the target language proper. At each stage, the text being processed appears either in a natural language, or in a more abstract but still language-specific coding. The actual work of language translation is done at a more abstract level than directly from source text to target text (which would consitute a *direct* system), and is thus more tractable and amenable to systematisation.

It is at least theoretically possible to have an MT system in which the abstract representa tions of the source text and target text are sufficiently removed from their corresponding languages as to be identical, and not to require a distinct transfer procedure. Such a system would be employing an *interlingua* strategy: the source text is rendered in an abstract interlingua, a representational system able to do underlie both the source and the target language, and from which the target language text could be directly generated.

The following ‘pyramid’ is frequently used to illustrate the difference between the three strategies. In the diagram, the horizontal axis represents bilingual transfer, while paths to wards and from the pyramid apex represent monolingual analysis and generation. As can be seen, transfer represents a compromise between direct and interlingua translation: it re duces the amount of bilingual transfer required with respect to direct translation, and it reduces the amount of analysis of the text required with respect to interlingua translation.



Both interlingua and transfer systems have their respective advantages. Interlingua systems are prima facie more attractive for multilingual MT systems: adding a new language to a system with *n* languages augments the number of modules in the system (analysis, generation, and transfer) by $2n$, whereas a transfer system requires $3n^2$ more modules. The mod ule design task thus should become forbiddingly large for transfer systems.

There are two major disadvantages with interlingua systems; these have meant that the in terlingua approach is used only in a minority of MT systems. The first disadvantage is the theoretical and practical difficulty of defining a consistent interlingua. Attempts to do so inevitably occur under the shadow of efforts in linguistic philosophy to define a universal, language-independent representation of meaning—ranging from classical logic, through the pasigraphies of the early modern era, to recent frameworks such as Schank’s Conceptual Dependency graphs and Wierzbicka’s Natural Semantic Metalanguage. This is a huge and open-ended endeavour, with major disagreements between the different theories proposed, and which cannot be expected to yield quick-and-dirty results of the kind probably re-

[1] See Hutchins & Somers(1992) Chapters 4 & 6 for further discussion of translation strategies in MT.

quired for a working interlingua system. MT researchers understandably find the task more trouble than it is worth.

The second disadvantage to the interlingua approach is that the process of rendering a natural language text into an abstract interlingua requires much more detailed lexical and structural analysis and disambiguation of the source text than is either desired or—frequently enough—feasible. In particular, a interlingua cannot exploit the 'free rides' used in transfer systems when a language pair share a categorisation of lexemic space or a common phrasal structure. For instance, an interlingua may contain a distinct lexeme for 'elder brother'and 'younger brother', to reflect the distinct lexemes used for these two notions in Japanese. (This follows from the requirement that the interlingua have all the information required to generate any target language text.) While the differentiation between the two is required for Japanese, though, it is not required between, say, French *frère* and English *brother*. But a French-English-Japanese interlingua system would have only one interlingua, by definition; translation between French and English in the system would thus mandate the disambiguation of 'older/younger brother', even though it serves no purpose for that particular task.

More generally, the further removed the target language is from the source language, the more forbidding the transfer task. In order to encompass all possible meaning transfers in the system, the interlingua needs to be at least as removed from the system languages as any system language is from any ot her. Depending on the nature of the interlingua used, it could be even further removed than this. As a result, the transfer task can become downright intractable; and in practice even interlingua efforts moderate their ambitions. Thus, Esperanto-based DLT made use of an existing (albeit artificial) human language as its interlingua, converting their system in effect into a double-direct system (Schubert 1988b). Rosetta, a system based on Montague grammar, forces interdependence of analysis and generation subgrammars (Hutchins & Somers 1992:292).

For these reasons, most multilingual systems developed to date (involving, for the most part, Indo-European languages, particularly in the context of the European Community) use transfer rather than interlinguas. It can be argued that, since the target languages are structurally close enough to each other in any case, the ability to exploit 'free rides' (structural similarities between two languages), and the well-defined nature of the analytic target (a known target natural language, rather than a nebulous interlingua yet to be worked out) outweigh the disadvantages of writing O(n^2) rather than O(n) translation modules for the system.

There are two observations worth making in this respect. First, there is a continuum ranging between direct and pure interlingua systems, as can be deduced from the pyramid dia gram (see also discussion in Hutchins & Somers (1992:128)). The target of source text analysis can be more or less abstract with respect to the source language—in terms of the diagram, closer or further from the pyramid apex. The same also holds for the generation side of the pyramid. Indeed, it is possible for an MT system to involve an interlingua in processing the linguistic structural aspects of the text, while using transfer for lexical content. What this means is that there is no intrinsic difference between a transfer abstract representation and an interlingua: the two may be formally quite similar—particularly if the transfer representation is highly abstract to begin with. Thus, transfer systems can draw on the results of interlingua systems and vice versa, in constructing their respective abstract representations of language.

The second observation is that in a typologically more diverse linguistic area than Europe—such as the Pacific Rim, in which an MT system would conceivably involve English, Japanese, Chinese, Malay/Indonesian, Thai, and Vietnamese, each belonging to a different language family—one would expect much fewer 'free rides'to be available to the MT system, as the language share fewer structural features and categorise lexical space more differently. Such an MT system is thus forced to do more work, either in analysis (further abstracting input) or in transfer, in order to generate a cceptable output. In that regard, an interlingua approach, or at least a much more abstract transfer approach, begin to look more attractive.

2. What an Interlingua Needs To Do

Given that an interlingua approach to MT is worth at least investigating in this region, it is worthwhile to examine what that interlingua should look like. This discussion takes place against the background of Schubert's (1988a) and Boitet's (1988) discussion of the merits of Esperanto as an MT interlingua.

There are two stages in the language transfer process: structural and lexical. As noted, it is possible for MT to handle the two stages distinctly, using an interlingua for language structure and transfer for the lexicon. The requirements on an interlingua as to its structure and its lexicon are different, and I consider them separately. Note however that, with regard to case grammar in particular, the two stages cannot be isolated from either. Thus, a structural interlingua will necessarily have a case grammar which may be distinct from that of the source or target language; a lexical transfer module would need to be aware of what choices the structural interlingua has made in analysis.

If an interlingua rather than transfer is chosen to handle the lexicon, one would expect the interlingua to be able to represent all the shades of meaning realisable as distinct lexemes in one of the target languages. It is commonly assumed that a formal-semantic system involving semantic primitives is the way to do this. However the DLT designers, in particular, have contended that no artificial language can be more explicit than the natural languages on which it is ultimately based, and that therefore the disambiguation involved in formulating semantic primitives is open-ended but pointless. They therefore have used Esperanto instead, as a regularised natural-like language.

So the two alternatives for a lexical interlingua are a natural-like interlingua and a formal system incorporating some notion of semantic primitives. The advantages claimed for the former are that the interlingua encompasses the expressiveness and flexibility of natural languages, denied to formal systems; that its vocabulary is already extant and standardised; and that it is human-learnable and readable (Hutchins & Somers 1992:299). A further advantage adduced in Sadler (1989) is that any world-knowledge information made available to the MT system (such as DLT's Bilingual Knowledge Bank) should be coded in the same language as what the system uses to represent meaning internally—that is, in the interlingua. Such a knowledge bank is easier to compile—and indeed, to assemble from existing materials—in a human language than in an artificial language fabricated *de novo*.

These conclusions are all arguable,however. If an interlingua is to minimise distortion in lexical translation, the major desideratum seems to be not so much that the interlingua be formal (Montagovian, propositional-logical, etc.), as that it be maximally explicit—so that the meaning mappings between source, interlingua and target language be as lossless as possible (or, where loss in transfer occurs, it is at least predictable from the interlingua definition, and can be compensated for.) It may well not be feasible to have the interlingua make all the lexical distinctions possible in any target language—for example, the interlingua might not, like Japanese, have a separate lexeme for *older brother* and *younger brother*. But if the lexemes of the interlingua are defined explicitly enough, composing such semantic units within the interlingua when required should not prove a problem; indeed, that is the approach taken by representational systems based on semantic primitives.

However, precisely because Esperanto is used just like any human language, its dictionaries look no different to natural language dictionaries (see for example the authoritative Esperanto dictionary, Waringhien, G. *et al.* (1987), which resembles the French *Larousse* dictionary fairly closely.) As a result, lexical space still needs to be made explicitly delimited enough to make such a language a functional interlingua; and this process now needs to proceed for the interlingua without the tools afforded by a formal approach. So Esperanto can prove just as problematic in lexical transfer as artificial interlinguas—if not more so. The segmentation of lexical space done by both kinds of interlingua is by necessity arbitrary (in the absence of a commonly agreed upon system of semantic primitives), and Esperanto lacks the degree of explicitness required to account for this arbitrariness in the MT task.

Lack of explicitness is an even greater problem for a natural-like interlingua used as the structural interlingua of the system. If the interlingua happens to be typologically similar to the target languages of the system, then all is well: the interlingua is not doing significantly more work than a transfer system would. Arguably, this is in fact the case for Esperanto and European languages, with respect to most aspects of the grammatical system if not morphology. But if this is not the case, then converting source to interlingua and interlingua to target structure becomes much more complex, and may offset any advantages gained by using an interlingua rather than transfer MT. The way to at least minimise this problem is to ensure the grammatical system of the interlingua is well-defined enough—using explicit mappings of semantics to surface structure rather than conventional, non-universal natural-language traits such as case and mood—that the structural interlingua does not founder on grammatical incompatibilities.

For example, nominative and accusative case may be quite appropriate for a European interlingua, but would yield no advantage over language-specific transfer in a more typologically diverse system, where the grammatical construct of subject or direct object may not be as relevant. A system with an explicit semantic case grammar, on the other hand, might be doing point less busy-work when translating between English and German, for which subjects usually correspond to subjects (although this is not always the case—cf.*I like it* with *Es gefällt mir.)* But when languages are involved in which the notion of subject is not as straightforward (as with Chinese and Japanese, in which subjecthood competes with topichood, or languages with applicative verb formation such as Austronesian languages, in which oblique case roles can be promoted to subject or direct object), an explicit case grammar probably ends up making the transfer task much simpler than case-by-case treatment of each language.

MT system implementers using natural-like interlinguas need also to be wary of becoming entangled with the grammatical idiosyncracies of their interlingua. In the case of DLT, the system designers went out of their way to consult with the Esperanto community, in order to min imise the divergence of human Esperanto and the DLT interlingua. Aside from the peculiar spectacle of the DLT designers finding the Esperanto grammar used in their questionnaire criticised (Silfer 1986), the premise that the regularity of an interlingua be compromised in order to meet the characteristics of its human model (for example, gender marking for third personal singular pronouns but not third person plural) seems decidedly odd. The designers'argument that the interlingua should be readable to the designers does not seem convincing here: the need to retrain a designer team of tens to recognise a few strategic alterations to Esperanto (assuming everyone in the team was already conversant in the language!) does not seem to outweigh the deterioration in system performance resulting from interlingua ambiguities.

The expressivity of the interlingua, another argument adduced in favour of natural-like interlinguas, is probably not an issue for the types of texts and applications an MT is currently expected to handle. There is certainly nothing more logical or computationally tractable about Esperanto idiomatic expressions and cultural referents than with their English or Japanese equivalents.[2] Finally, the amount of text in Esperanto that could serve as a corpus on which a Knowledge Base could be constructed is limited in comparison to such text in English, particularly with regard to technical fields. So while some saving of time might be achieved over an artificial interlingua in compiling such a corpus, the saving would not be as significant as one would think; and MT designers would still need to work out some way of getting a large amount of source language information into an interlingua database—whether the interlingua is formal or natural-like.

By the same token, however, not all aspects of human language are equally amenable to formalisation—or at least the manners of formalisation we are familiar with to date in computational linguistics. Problem areas include what Hallidayan linguistics has termed the *interpersonal* and *textual* usages of language: that is, the uses of language to express attitude, social aspects of language, and textual cohesion (as opposed to the *ideational* usage of language, to express events in the real world.) These are all areas in which text generation has been traditionally deficient (notwithstanding pioneering work such as Hovy (1988).)

While the types of text expected to be handled by MT do belong to a restricted genre with a limited repertoire of stylistics, improving quality of style nevertheless remains a major goal not only of MT, but of text generation in general. The quality of text will be improved significantly once the textual metafunction of language is taken into account in systems—that is, once text cohesion is successfully approached. Stylistics is also a factor that needs to be taken under account in higher-quality MT; if the MT system can appreciate from the outset the distinction between stylistic variants of what is the same concept as far as formal semantics is concerned, they will become able to emulate the requisite style to at least some extent in output, reducing somewhat the requisite amount of post-editing.

One may conclude that, if an MT interlingua is to be tenable, it needs to be not only autonomous and expressive, as Schubert (1988b) has argued, but also explicit. Indeed explicitness (in representation of not only ideational semantics, but also interpersonal and textual meaning) would probably compensate for whatever deficiencies a formal interlingua has in autonomy (dependence on natural languages in its definition) and expressiveness. If the definition of the formal interlingua is explicit, then its dependency on a natural-language metalanguage can be factored in to its use in MT. By contrast, capturing Esperanto lexical semantics is probably as slippery an enterprise as capturing natural language semantics—and does not seem to constitute a secure basis for an internal unambiguous representation of meaning. As for expressiveness, an explicit system—even if not so formal as to attempt a componential analysis of interpersonal meaning—would at least give something tangible for

[2] One should bear in mind that the very reason Esperanto has become as expressive as natural human languages is that it has picked up natural-language ways of saying things. The role of German (the prestige language for the first generation of Esperanto-speakers) has been crucial in this respect.

the MT system to work with. By contrast, Esperanto word order may be expressive, but extracting interpersonal meaning from marked Esperanto word order is just as tricky as in German (for example, in an Esperanto—or German—OVS sentence, is O the focus, or S the topic?)

So an MT interlingua—particularly in a typological diverse context justifying the use of an abstract interlingua rather than transfer—would probably have to look something like the formal systems we are accustomed to, but embellished with features allowing interpersonal and textual meaning to be handled. It probably should not look like Esperanto, although admittedly the difference in performance between the two types of interlingua systems would probably not exceed an order of magnitude.[3] Thus, it would encompass a case grammar, a system of semantic primitives and lexical compositionality, some form of predicate logic (formal or informal), and an unambiguous syntax (whether it also needs a denotational semantics is debatable); but it would also include at least some labels indicating topicality, information flow, attitutde, hedges, and register. In essence, it would be a predicate logic with non-ideational graffiti.

3. Lojban

Lojban, I suggest, is such a representational system, and shows what such a language would look like overall, although the details of the system may still be subject to debate.

Lojban (Cowan *et al.* 1996) is an artificial language designed collectively by the members of the Logical Language Group, as a continuation of the earlier Loglan project (Brown 1960).[4] Loglan was originally intended as a test of the Sapir-Whorf hypothesis, according to which there is a correlation between the language spoken by a people and their world-view. It was reasoned that, if a spoken language were to incorporate the metalanguage of modern formal logic, this language would differ significantly enough from natural languages that the world-view of speakers of this language would differ measurably from those of non-speakers, if the Sapir-Whorf hypothesis holds.

Needless to say, the practicalities and ethical issues involved in building up a native-speaker Loglan community are considerable, and the anecdotal effcts reported from casual second-language use of the language do not count for much. At any rate, the Sap ir-Whorf hypothesis is no longer the primary motivation for interest in Lojban amongst supporters of the language (nowadays drawn to a great extent from the Internet.) Rather, the major attraction Lojban holds is as a quasi-formal model of language, based on predicate logic, which attempts to cover the functionality of human language as broadly as possible. A good deal of this interest focusses on the possible computational applications of Lojban *qua* formal model of language. Thus, Lojban has already been proposed as a machine translation interlingua to the USA Department of Defence by the Logical Language Group (Le Chevalier 1992).

Although professional linguists have been involved with the language design effort, the language has been designed for the most part by amateur linguists. As a result, it is possible to take issue with particular facets of the design, and I certainly do not contend that Lojban should be taken on as is, without question, as an MT interlingua.[5] Rather, I would argue that an interlingua of the type I have discussed (‘a predicate logic with non-ideational graffiti’) would resemble Lojban greatly, and designers of such interlinguas would profit from drawing on the results of Lojban design.

Furthermore, the fact that this is an amateur effort may well make it attractive for MT researchers. MT requires broad linguistic coverage for an interlingua—it needs to represent all the language semantics involved in meaning transfer between language. An amateur

[3] I should point out that I do not intend to attack the applicability of Esperanto for the purpose for which it was designed—a function it fulfils admirably—but its usefulness as an MT interlingua. As has been admitted by one of the DLT designers (Maxwell 1992), Esperanto was adopted *faute de mieux,* and more formal interlinguas could not be ruled out as useless to the task.

[4] The Lojban group broke off from Dr Brown’s group in the mid ’80s, as a result of a copyright dispute on the grammar and vocabulary of the language. The Logical Language Group, unlike Brown’s Loglan Institute, have kept the grammar and vocabulary of their variant of the language in the public domain.

[5] Nevertheless, a perusal of Cowan *et al.* (1996) makes it clear that a good deal of thought has gone into the project; debate about theoretical issues in the language design continues to take place on the Lojban mailing list (`lojban@cuvmb.cc.columbia.edu`), although the language design is now essentially complete.

effort like Lojban, with the avowed intention of producing a speakable language, is much likelier to cover language breadth well, even if does not explore semantic depth as soundly as a professional linguistic research project. Linguistic research, by contrast, concentrates on specific problems of meaning and form; as a result, they do not typically attempt to cover the whole of language, but only those parts of language relevant to the particular research topic. One would therefore look in vain for accounts of text cohesion in Montague grammar, or case grammar in Rhetorical Structure Theory. While the ‘toy grammars’ described by such accounts of language can be highly illuminating, they do not in isolation do what an interlingua needs to do.

For an interlingua to do its job properly on several linguistic levels, it needs to be eclectic in combining insights from different theories; the heterogeneity of the theories involved probably mean the interlingua designer needs to sacrifice elegance in favour of comprehensiveness.[6] The resulting interlingua would probably not differ essentially from Lojban—although as I argue below, different choices might be made in design particulars.

The major design features of Lojban which make it of interest as an interlingua are as follows:

- A set of some 1350 morphologically basic predicates (`gismu`), each with a well-defined argument structure (that is, a specification of the number of argument places for each predicate, and of what arguments fit in each place.) The number of predicates is rather expansive in comparison to the typical semantic primitive sets proposed (which rarely exceed 100.) As a result, Lojbanists tend not to refer to the `gismu` as semantic primitives. The emphasis in the design has been on comprehensiveness and practicality, rather than minimalism. The predicates are drawn from basic instructional vocabulary and word frequency lists compiled for various languages; loan words can also be incorporated into the language.

It is fair to say that the Lojban set of basic predicates is arbitrary; but by erring on the side of inclusivity, it becomes much easier to ensure that the predicates, either on their own or in compositional combination, cover a reasonable breadth of lexical space in a practical manner. A set of 100 true semantic primitives would enable more thorough definitions of semantically complex terms; but the paragraph-length renderings that can result and are so typical of e.g. Natural Semantic Metalanguage are not a viable alternative in the kind of MT lexical transfer we can envisage in the near future. Furthermore, while cultural bias may creep into some of the more semantically complex Lojban predicates (such as `vrude` ‘virtuous’), thereby reducing the autonomy of the interlingua, the system always has recourse to semantically simpler predicates, which can be used to build up a less covert representation of the intended notion as in NSM. A workable, practical set of basic primitives is therefore preferrable to a smaller and more rigorous set, which would result in complex lexical transfer. The explicitness of the predicate argument definitions in Lojban helps ensure that the larger set does not become vague and unmanageable. The fact that no comparable set of predicates exists defined in as much detail and so comprehensive should make Lojban’s `gismu` set of interest at least as raw material for an interlingua.

The formal logic orientation of Lojban, however, forces a particular structure on these predicates which might be problematic for MT. In particular, Lojban semantic roles are primarily structured by predicate definitions, rather than a predicate-independent case grammar. Lojban has the capacity to use a case grammar (case grammar labels are included in the Lojban equivalents of prepositions, which add adjuncts to predications), but it is the place structure of individual predicates that is taken as definitional. For example, the first argument (x1) of the predicate for ‘angry’(`fengu`) and the second argument of ‘severe’ (`jursa`) are both experiencers of the predicate event; but this commonality is not reflected in the surface form of Lojban sentences, and has to be teased out of the definition of `fengu` and `jursa`. An interlingua may well want to accord more significance to a case grammar in exploiting languages’ tendency to give the same morphological treatment to

[6] In a roundabout way, this accords with Wilks’(1992:279) reaction to the success of SYSTRAN: “There is no theory of language structure so ill-founded that it cannot be the basis for some successful MT.” Wilks (1992) gives thorough arguments against the usefulness of formal logic in AI, and MT in particular; as far as Lojban is concerned, I would contend that the main value in Lojban as an interlingua, and a representational system in general, lies not so much in its use of formal logic structures, as in the explicitness of definition these structures afford. It is this explicitness, rather than any commitment to model-theoretical semantics, which makes Lojban a good candidate interlingua.

arguments in the same semantic case. An interlingua version of Lojban might therefore elevate case grammar to a definitional role with respect to the basic predicates.

A second problem inherent in the formal orientation of Lojban is the categorial way it treats the segmentation of lexical space. This makes less classical approaches to semantics, such as prototype theory, less straightforward to express. For example, the definition of the predicate `botpi` ‘bottle’ is the set of all objects x1 which are containers for substance x2, made of material x3, and having a lid x4. For the predicate `botpi` to hold, all four entities must exist; this means that `botpi` is by default not a fuzzy-logical predicate, and cannot be used to denote ‘sort-of bottles’ (bottle-shaped objects which do not act as containers; lidless bottles; bottle shapes; etc.) Lojban can only deal with such entities by periphrasis (expressions such as `botpi simsa` ‘bottle similar’.) This problem can be circumvented within human-like interlinguas, which would use the same metaphorical and metonymic semantic extensions of lexemes as would target languages; but one should be wary that not all such extensions are cross -linguistically valid. It is not difficult to embed ‘fuzzy-logical graffiti’(hedging expressions) within Lojban predications to make them compatible with prototype theory; but it is open to question how pertinent this issue is with the types of texts likely to be subject to MT anyway.

- A well defined syntax, already implemented in YACC and consisting of some 600 rules. (The Extended-BNF equivalent syntax has some 80 rules.) Unambiguous constituency is guaranteed by using terminators to designate constituent boundaries; for example:

`mi`	`jarco`	`le`	`nanmu`	`poi`	`ke’a`	`citka`	`le`	`plise`	`ku’o`
I	show	the	man	REL	HEAD	eat	the	apple	END-REL

I showed the man {who was eating the apple}

`mi`	`jarco`	`le`	`nanmu`	`poi`	`ke’a`	`citka`	`ku’o`	`le`	`plise`
I	show	the	man	REL	HEAD	eat	END-REL	the	apple

I showed the man {who was eating} the apple

A lack of syntactic ambiguity is one of the essential prerequisites of any interlingua. It does not mean, of course, that the task of resolving syntactic ambiguity in the source language becomes any easier; nor is the question of whether a syntactically unambiguous language is actually speakable relevant here. The adoption of an interlingua means that, in general, the system forfiets the capacity to exploit ‘free rides’ in the presence of the same syntactic ambiguity in both source and target language; but as already argued, there are much fewer such ‘free rides’ to be exploited in a typologically diverse MT system. (The notorious example ‘The man saw the girl with the telescope’, for example, does not yield a free ride on translation into Japanese.)

- A compositional semantics of word compounding—allowing the predicate structure of compound predicates to be derived from the predicate structures of constituent predicates. For instance, given the predicates `nenri` ‘x1 is inside x2’ and `klama` ‘x1 goes to x2 from x3 via x4 using means x5’, one can derive not only the compound predicate `nenryklama` ‘to inside-go=to enter’, but also its arguments: ‘x1 is inside x2, and goes to x2 from x3 etc.’

In Lojban’s model-theoretical view of semantics, the ability to derive compound predicate structures is tantamount to defining the semantics of compound predicates. This means that the interlingua lexicon can be expanded in a principled manner, without any loss in semantic explicitness. It also means novel concepts can be expressed in the interlingua lexically rather than phrasally—which should make transfer from source lexeme to target lexeme much easier than attempting to map source lexeme to target lexeme via interlingua phrase (a problem with representations of meaning like NSM, as noted above.) Word compounding semantics is a recent addition to the design of Lojban (in which I have been involved); in view of the intended human use of Lojban, and the influence of natural-language word-compounding semantics, Lojban compounding curren tly allows for ambiguity in the head-modifier relation, although reducing it to three or four possible alternatives based on predicate arguments shared between the head and modifier; there is also some flexibility allowed in applying the compounding rules. An interlingua making use of such compounding would presumably eliminate such ambiguities by morphemically tagging the head-modifier relation chosen.

- A system of particles (‘UI words’) lying outside the predicate logic of Lojban’s meta-language, and used as attitudinals, evidentials, discourse markers, and information flow markers—in other words, performing the work of interpersonal and textual meaning, not done by the predicate core of Lojban grammar. For instance:

le	nu	mi	jai	rinka	loi	nu	darlu	cu
the	EVENT	I	RAISING	cause	some-MASS	EVENT	argue	PRED

milxe	zo'o	lakne
mild	HUMOUR	possible

There is a slight possibility :-) I will stir up some controversy

zu'unai	lo	bi'u	cnino	tadji	cu	fapro
on the other hand	a	NEW-INFO	new	method	PRED	oppose

le	di'u	nabmi
the	previous-mention	problem

On the other hand, there is an improved method to combat this problem

It is through its system of interpersonal and textual markers that Lojban meets the interlingua requirement of 'non-ideational graffiti' I have already mentioned. As was the case with basic predicates, the system of modifiers elaborated within Lojban emphasises inclusiveness rather than rigour, and individual choices and omissions may be debated in formulating the particulars of the system interlingua, according to the requirements of the source and target languages and the types of texts processed by the MT using the interlingua. What matters, however, is that a system of particles can be set up outside the predicate semantics of the representational system, modifying terms within the predicate semantics, and representing a level of language function typically omitted from MT and text generation systems as unformalisable. Inasmuch as these levels of language function really are unformalisable—that is, not expressible in terms of predicate logic or other representations of ideational meaning—the appropriate way to handle them is to devise labels with at least some measure of cross-linguistic validity, use them to modify the appropriate constituents within the interlingua representation of the text, and confide the work of realisation to the generation module.

We have already seen that it is possible for a symbol to use interlingua as to structure, but transfer as to lexicon. Non-ideational meaning constitutes a third language level, for which an MT system may behave modularly. It is not impossible that transfer, rather than interlingua, will be necessary for the task—that it will prove too difficult to formulate language-independent labels for non-ideational meaning, or that linguistic diversity in non-ideational meaning is so great and idiosyncratic that the language-pair approach of transfer MT proves more perspicuous. A framework similar to Lojban's could none the less provide the basis for the abstract transfer representation of non-ideational meaning, and the manner in which it is woven into the ideational text representation.

4. Conclusion

Lojban is a representational system which has some potential in meeting the desiderata of an MT interlingua. Lojban by no means makes any easier the problems involved in the analysis of source texts or the selection of target text structures. Indeed, being much more structurally abstract than the source and target languages, an interlingua like Lojban makes the text-analytic task much more complex, forcing upon the system a much greater number of ambiguities to contend with. I maintain, however, that a linguistic environment such as that of the Pacific Rim, in which a multilingual MT system has to cope with a greater typological variety of languages than in Europe, increased abstraction in the language transfer task (either with interlinguas proper, or at the least with much more abstract transfer representations—or admixtures of the two, depending on the particular transfer task) are inevitable if the MT task is to remain practical. For the same reasons of practicality, such a representational system should be characterised by breadth rather than depth of coverage, and by comprehensiveness rather than rigour of design. If the representational system is to have those characteristics, yet remain identifiably close to the definitional machinery of formal semantics, then I do not think it will differ in essence from Lojban; and designers of such representational systems would do well to take Lojban under consideration.

References

Blanke, W. 1988. Terminologia Esperanto-Centro: Efforts for Terminological Standardization in the Planned Language. In Maxwell, D., Schubert, K. & Witkam, T. (eds.) *New Directions in Machine Translation: Conference Proceedings, Budapest 18–19 August, 1988.* (Distributed Language Translation Series) Dordrecht (Netherlands): Foris. 183–193.

Boitet, C. 1988. Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation. In Maxwell, D., Schubert, K. & Witkam, T. (eds.) *New Directions in Machine Translation: Conference Proceedings, Budapest 18–19 August, 1988.* (Distributed Language Translation Series) Dordrecht (Netherlands): Foris. 93–108.

Brown, J.C. 1960. Loglan. *Scientific American* **202:6**. 5363.

Cowan, J. *et al.* 1996. *Draft Lojban Reference Grammar.* Electronically available from `ftp://powered.cs.yale.edu/pub/lojban/draft/refgrammar.`

Hovy, E.H. 1988. *Generating Natural Language Under Pragmatic Constraints.* PhD thesis, Yale University.

Hutchins, W.J. 1986. *Machine Translation: Past, Present, Future.* Chichester: Ellis Horwood.

Hutchins, W.J. & Somers, H.L. 1992. *An Introduction to Machine Translation.* London: Academic Press.

Le Chevalier, R. 1992. la lojbangirz's First Research Proposal. *ju'i lobypli* **16**. 27–31.

Maxwell, D. 1992. DLT—Esperanto-based Machine Translation. *ju'i lobypli* **16**. 34–35.

Morneau, R. & Cowan, J. 1993. Criticisms of Lojban as a Tool for Machine Translation. *ju'i lobypli* **17**. 15–21.

Sadler, V. 1989. *Working with Analogical Semantics: Disambiguation Techniques in DLT.* (Distrubuted Language Translation 5). Dordrecht (Netherlands): Foris.

Schubert, K. 1988a. Att knyta Nordens språk till et mångspråkigt datoröversättningssystem. In *Nordiske Datalingvistikdage og symposium for datamatstøttet leksikografi og terminologi 1987, Proceedings.* Copenhagen: Institut for Datalingvistik, Handelshøjskolen i København. 204–216. (Cited in Blanke 1988)

Schubert, K. 1988b. The Architecture of DLT—Interlingual or Double Direct? In Maxwell, D., Schubert, K. & Witkam, T. (eds.) *New Directions in Machine Translation: Conference Proceedings, Budapest 18–19 August, 1988.* (Distributed Language Translation Series) Dordrecht (Netherlands): Foris. 131–144.

Silfer, G. 1986. DLT-Projekto: La Gvidantoj Enketas (The DLT Project: The Administrators' Survey). *Planlingvistiko* **17**. 1–6.

Waringhien, G. *et al.* 1987. *Plena Ilustrita Vortaro de Esperanto kun Suplemento.* 3rd ed. Paris: SAT.

Wilks, Y. 1992. Form and Content in Semantics. In Rosner, M. & Johnson, R. (eds.): *Computational Linguistics and Formal Semantics.* (Studies in Natural Language Processing) Cambridge: Cambridge University Press.