AD-A269 723

(12)

APPROXIMATING AN INTERLINGUA IN A PRINCIPLED WAY

Eduard H. Hovy
4676 Admiralty Way
Marina del Rey, CA 90292
Sergie Nirenburg
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213

February 1993
ISI/RR-93-345

DTIC
ELECTE
SEP 21 1993
S A D

This document has been approved for public release and sale; its distribution is unlimited.

In Proceedings of the DARPA Speech and Natural Language Workshop, Hawthorne, NY, February 1992

93-21769
14pg

93 9 17 053

REPORT DOCUMENTATION PAGE

FORM APPROVED
OMB NO. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimated or any other aspect of this collection of information, including suggestings for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	February 1992	Research Report

4. TITLE AND SUBTITLE
Approximating An Interlingua In a Principled Way

5. FUNDING NUMBERS
MDA903-87-C-641

6. AUTHOR(S)
Eduard Hovy and Sergei Nirenburg

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
USC INFORMATION SCIENCES INSTITUTE
4676 ADMIRALTY WAY
MARINA DEL REY, CA 90292-6695

8. PERFORMING ORGANIZATON REPORT NUMBER
RR-345

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
ARPA
3701 Fairfax Drive
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
In Proceedings of the DARPA Speech and Natural Language Workshop, Hawthorne, NY, February 1992

12A. DISTRIBUTION/AVAILABILITY STATEMENT
UNCLASSIFIED/UNLIMITED

12B. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We address the problem of constructing in a principled way an ontology of terms to be used in an interlingua for machine translation. Given our belief that a true language neutral ontology of terms can only be approached asymptotically, the construction method outlined involves a stepwise folding in of one language at a time. This is effected in three steps: first building for each language taxonomy of the linguistic generalizations required to analyze and generate that language, then organizing the domain entities in terms of that taxonomy, and finally merging the result with the existing interlingua ontology in a well-defined way. This methodology is based not on intuitive grounds about what is and is not 'true' about the world, which is a question of language-independence, but instead on practical concerns, namely what information the analysis and generation programs require in order to perform their tasks, a question of language-neutrality. After each merging is complete, the resulting taxonomy contains, declaratively and explicitly represented, those distinctions required to control the analysis and generation of the linguistic phenomena. The paper is based on current work of the PANGLOSS MT project.

14. SUBJECT TERMS
Machine Translation, Interlingua, semantic ontology, computational linguistics

15. NUMBER OF PAGES
6

16. PRICE CODE

17. SECURITY CLASSIFICTION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reoprts. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to stay within the lines to meet optical scanning requirements.

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month,a nd year, if available (e.g. 1 jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element numbers(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	PR	- Project
G	- Grant	TA	- Task
PE	- Program Element	WU	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the repor.

Block 9. Sponsoring/Monitoring Agency Names(s) and Address(es). Self-explanatory

Block 10. Sponsoring/Monitoring Agency Report Number. (If known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of ...; To be published in... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

- DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
- DOE - See authorities.
- NASA - See Handbook NHB 2200.2.
- NTIS - Leave blank.

Block 12b. Distribution Code.

- DOD - Leave blank.
- DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
- NASA - Leave blank.
- NTIS - Leave blank.

Block 13. Abstract. Include a brief (Maximum 200 words) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (NTIS only).

Blocks 17.-19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contins classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

APPROXIMATING AN INTERLINGUA IN A PRINCIPLED WAY

Accesion For
NTIS CRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

Eduard Hovy[1] and Sergei Nirenburg[2]

1) Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695

2) Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213

DTIC QUALITY INSPECTED 3

Abstract

We address the problem of constructing in a principled way an ontology of terms to be used in an interlingua for machine translation. Given our belief that a true language-neutral ontology of terms can only be approached asymptotically, the construction method outlined involves a stepwise folding in of one language at a time. This is effected in three steps: first building for each language a taxonomy of the linguistic generalizations required to analyze and generate that language, then organizing the domain entities in terms of that taxonomy, and finally merging the result with the existing interlingua ontology in a well-defined way. This methodology is based not on intuitive grounds about what is and is not 'true' about the world, which is a question of language-independence, but instead on practical concerns, namely what information the analysis and generation programs require in order to perform their tasks, a question of language-neutrality. After each merging is complete, the resulting taxonomy contains, declaratively and explicitly represented, those distinctions required to control the analysis and generation of the linguistic phenomena. The paper is based on current work of the PANGLOSS MT project.

1 The Problem with Interlinguas

This paper presents a method of constructing in a principled way an ontology of terms to be used in an interlingua for machine translation. The method involves taxonomizing the linguistic phenomena apparent in each language and then merging the taxonomy with the interlingua ontology to produce an ontology that explicitly records the phenomena that must be handled by any parser or generator and is neutral with respect to the languages handled by the system.

1.1 What is an Interlingua?

In interlingual machine translation, the representational power of the interlingua is central to the success of the translation. By interlingua we mean a notation used in MT systems to represent the propositional and pragmatic meaning of input texts; an *interlingua text* that represents the meaning of a source language text is produced by computational analysis, and is then used as input to a generation module which realizes this meaning in a target language text.

An interlingua consists of the following three parts:

- a collection of terms: the elements that represent individual meanings (of lexical items, pragmatic aspects, etc.); this collection is organized in a multiply interconnected semantic network;
- notation: the syntax to which well-formed interlingua texts conform;
- substrate: the knowledge representation system in which interlingua texts are instantiated and which provided the infrastructure for reasoning with the knowledge encoded in interlingua texts.

The term network can be thought of as a conceptual lexicon of basic meaning distinctions that represent entities in the world. While it is not feasible to produce a complete model of the world, it is possible to produce a well-formed subset in which each term's definition contains enough features to differentiate it from all other terms in the subworld. A good taxonomization enables the sharing and inheritance of properties and facilitates definitional brevity and expressive power. Any given self-contained subworld includes a set of terms which it can share with other subworlds. These are to be found at higher levels of the network, closer to its root node. It is customary to call a subnetwork of terms that are expected to be present in most (if not all) subworlds an *ontology*, while the rest of a particular subworld is often called a *domain model*. Though the boundary between an ontology and a set of domain models is rather difficult to establish in the abstract, in practice, for any given set of domain models, it is straightforward to detect the shared higher levels.

The main purpose of an ontology and a domain model in MT is to support lexical-semantic analysis and generation. The open-class lexical units of a language have their meanings explained using terms from the ontology and the domain model. Connections are made in the lexicons for each particular language. (For a detailed description of one such interlingua-ontology-lexicon combination see [Nirenburg & DeFrise 92, Meyer et al. 90], [Carlson & Nirenburg 90]).

1.2 Language Independence vs. Language Neutrality

By definition, an interlingua is language-independent and depicts the world directly, not through the medium of a natural language. Of course, particular interlinguas vary in their quality and utility. The two major properties of an interlingua which influence its quality are its expressive power (that is, the ability to support a different structure for each of a set

of ambiguous natural language expressions) and its capability to scale up without a major overhaul of its syntax (a point which relates directly to the task of term acquisition).

In theory, interlingua is language-independent. That is, all the languages handled are pairwise independent with respect to the interlingua, and the interlingua contains no "traces" of particular semantic distinctions predominant in a particular language or group of languages. The degree of language independence attained by an interlingua in reality is a matter of judgment.

The feasibility of language-independent representations has been debated many times. Since interlinguas are by definition language-independent, the debate revolves about the degree of overlap with language-specific knowledge and the status of such overlap. This debate is interesting but not practically relevant in the area of MT. A fully language-independent ontology is a worthy ideal. Since the acquisition-related tasks of selection, definition, and taxonomization of terms for an ontology and domain model are always performed by people, they are always open to being influenced by the distinctions made in the acquirers' languages as well as their "folk" models of the world.[1]

However, our approach to the problem of interlingua construction is more practical. We are concerned with the matter of constructing an interlingua that is language-neutral, i.e., one that enables analyzers and generators to operate optimally in a machine translation system. Under this approach it is still possible to maintain the central architectural tenet of interlingual MT systems — that it is possible to translate from a text to its interlingua text without regard to any eventual target language, and to translate from the interlingua text to a target without regard to the original source language.

We focus in this paper on the ontology, leaving aside questions of notation and substrate. Given the complexity of the ontology construction task, we believe that any given ontology will inevitably be closer to some languages than to others, both in the terms selected and their connections. We don't consider this to be a problem, as we would be able to augment the ontology and domain model as necessary whenever new domain models and new languages are added to the system.

What we do consider a problem is the lack of a generally established well-defined method for constructing ontologies. Most ontologies and domain models to date have been assembled based primarily on introspection, and often reflect the idiosyncrasies of the builders more than the requirements of the application (such as MT). Lacking well-founded guiding principles, the ontology builder is working in the dark. In this paper, we suggest that a better methodology uses data-oriented methods and linguistic knowledge to guide introspection.

[1] Consider, for example, the problem of representing colors: how many are there, and how are they organized? Even the basic distinction into the three primary colors so natural to any westerner is unlikely to occur to the Dani people of New Guinea, who have only two color terms (*mili* = dark-cool; *mola* = light-warm) and have proven difficulty in learning names for nonfocal colors [[Rosch 73].

2 A Method of Ontology Construction

2.1 Basic Criterion of Taxonomization

Recognizing that we cannot build a satisfactorily language-neutral ontology directly, we formulate an incremental procedure that, we believe, allows the acquirer to approach an optimal ontology asymptotically.

To develop the methodology we seek theoretically motivated answers to the following basic questions underlying ontology construction:

- What terms must be included in the ontology?
- How should the terms be organized in it?
- What level of detail should it reach?
- How closely can the ontology terms parallel the particular words and phrases of any single language?

In answering these questions, we exploit the fact that the ontology is inevitably going to be more or less language-specific to further our overall goal: a powerful interlingual MT system. The "closer" the ontology is to the source and target languages, the easier the process of ontology term acquisition and organization, and the less work the parser and generator have to do to bridge the gap between interlingua texts and source and target language texts. To find the point of "minimal distance" from all source and target languages, weighted as discussed in Section 2.5, we formulate the:

> **Basic criterion underlying terminology creation and taxonomization:** *The ontology must be just sufficiently powerful to represent the distinctions required to analyze the source languages and generate the target languages in the given domains.*

To define the distinctions needed to support each language, we start with a list of linguistic phenomena to be covered in it. This information can be extracted from a sufficiently rich grammar of the language. For example, for English the fact that nouns pattern into mass and count, the fact that adjectives and adverbs pattern differently than do verbs, or the fact that many different forms of possession (to have an arm, to have a spouse, to have a car) are expressed similarly. We then create a taxonomy of these linguistic abstractions, guided by their interrelationships. This taxonomy would, for these examples, contain nodes for *UncountableObj* and *CountableObj* under *Obj* to help handle nouns, nodes *Quality* and *Process* as high-level taxonomic organizers of adjectival/adverbial modifiers and verbal actions, and *GeneralizedPossession* as a high-level organizational point for the various types of possession. Any proposed term is accepted in the taxonomy only if it captures a distinct linguistic phenomenon of the language not otherwise handled. We end up with a "general" taxonomy of abstract entities that capture useful groupings of processing-oriented features.

Having established such a taxonomy, we then list all the entities that appear in the domain being addressed by the MT system (objects, actions, states, relations, plans, scripts

(action chains), etc.). What we are after is a taxonomization of these entities in such as way as to facilitate MT. Of course, the simplest scheme is no taxonomization at all: a list of the domain entities without any higher-level organizing entities. Then however each entity must contain enough information by itself to enable successful manipulation by the analyzer and generator (for example, *Company* must be explicitly annotated as being a *SocialOrg*, if this fact is used by analyzer and/or generator). A better solution is to group entities with the same features together and define nodes that represent generalizations. But which generalizations? We argue that most useful are those generalizations that play a role in the processing, analysis and generation (those that serve some differentiating function in the construction of an analysis or a clause) — exactly those contained in the taxonomy of linguistic phenomena. This taxonomy partitions the entities of the domain in ways which facilitate their treatment by the system.[2]

Thus, we use the linguistically inspired taxonomy as ontology under which to categorize all the domain terms (though this is not the only way to employ the language-oriented taxonomy as a guide, it is the most straightforward). By such a taxonomization, each domain term is located so as to inherit precisely the representational distinctions required to support its generation in the target language. Such taxonomies are fairly common in symbolically based NLP systems; see for example the ones used in JANUS (BBN) [Hinrichs et al. 87], LILOG (IBM Germany) [Lang 91], and Naive Semantics [Dahlgren et al. 90].

We call the resulting taxonomy, which consists of a linguistically inspired ontology and a domain-oriented domain model, the Language Base (LB) of the language. Since the resulting knowledge base is language-based, it is not necessarily best suited to support specialized forms of reasoning such as naive physics, legal reasoning, etc. But since our intention is MT, this poses no problem. An example of the abstractions that can be used as high-level organization for English is the Penman Upper Model [Bateman et al. 89], a taxonomy that has been used to support the generation of English text in several domains.

2.2 Toward Language Neutrality

It will be noted that the example terms employed above are semantic rather than syntactic (e.g., *Process* rather than *Verb*). Where obvious semantic correlates for syntactic generalizations exist, the strategy is to employ them, for they are more likely to appear also in the LBs of other languages. However, the result is still (at best) a mixture of semantic and purely syntactic generalizations, including generalizations for phenomena without any obvious semantic basis; for example, in English, the fact that verbs like "fill" and "spray" pattern similarly: "he filled the cart with hay" and "he sprayed the wall with paint" can be similarly transformed to "he used hay to fill the cart" and "he used paint to spray the wall".

[2] It must be understood that we are not stating that one must (or even can) perform exhaustive decomposition of all the facets of meaning of a concept to place it in the ontology. The basic criterion argues that just enough must be represented to enable the generation of the languages in question. Thus for example we do not need a term *Pink* for *Pale Red* unless the domain is such that there are implicative effects to saying "pale red" instead of "pink" (the former being a noncentral meaning in some contexts; see [McCawley 78] and [Jackendoff 85], p. 115).

Using abstractions from any particular language means giving up the basic goal of language neutrality. Since, however, we expect that any ontology is going to be slanted toward some language(s) more than others, we are willing to accommodate such "syntactic pollution" as long as it does not hamper MT. We next outline a method of progressively making a taxonomy more language-neutral to the point where all troublesome syntactic terms are removed. This process involves merging the LB for a second language with the LB for the first to form a hybrid taxonomy (also called a polylingua in [Kay et al. 91]), which we call the Ontology Base (OB), and then repeating the process for additional languages, according to the following procedure:

1. Construct the LB taxonomy for language 1. This is the ontology base (OB).
2. For each subsequent language,
   (a) construct the LB for its phenomena;
   (b) merge it with the existing OB, starting from the topmost entity and proceeding downward, as described in Section 2.3, ensuring that the lower-level, domain-specific, OB terms remain properly taxonomized.

The merging process can be considerably simplified if during construction of the LB for a new language (step 2(a)), the classes of the existing OB are used as a guide whenever various taxonomizations are possible.

From the languages we have examined, the following trend is apparent: the upper regions (ontologies) of the LBs are identical or nearly so. Differences occur in the middle regions, since they reflect language-particular information; the degree of cross-LB commensurability depends roughly on the closeness of the grammars. The lower regions are essentially identical for fairly international domains such as banking, science, technology, etc. — after all, however something like *Metal* is treated in the language, its descendants *Gold, Silver*, etc., are treated similarly and thus taxonomize together similarly.

As discussed in [Whorf 56] and [Lakoff 87], conceptual systems (of which OBs and LBs are an example) can be commensurate in several different ways. Lakoff lists five types of commensurability of two conceptual systems (p. 322): translation (a sentence-by-sentence translation preserving truth conditions is possible), understanding (a person can understand both alternatives), use (the same concepts are used the same ways), framing (situations are "framed" the same way and there is a one-one correspondence between the systems, frame by frame), and organization (when the same concept occurs in both, it is organized the same way relative to others that occur in both). Lakoff provides some examples: the systems of aspect in English and Hopi are commensurate by translation (since Whorf did translate sentences from one to the other) but not by use, since as Whorf's examples make clear (pp. 57–64), the "same" concepts for time are used very differently; the systems of Spatial Location in English and Mixtec are commensurate by translation but not by organization [Brugman 83], since for example sentences expressing the English meaning "on" translate to widely different Mixtec expressions depending on the shape of the lower object, among other things.

2.3 Merging Ontologies

Given the OB and an LB for a new language, we start the merging process from the topmost item(s) of the hierarchies. Usually, since the "meaning" of each item is captured by its interrelationships and ancestry, it is instructive to consider groups of closely related items simultaneously. The merging process involves one of three alternative operations for each item in the LB:

Identity: the LB item is identical to a corresponding item of the OB; they represent the same phenomenon, such as *DecomposableObj*. In this case, no further work is required beyond a name change to the OB item name. Identity may be difficult to determine for non-semantic items (one reason we suggest using semantic items when possible), or even occasionally for semantic ones, when their subordinate items differ. We discuss this point later in this section. However, in practise, identity of OB and LB items is common for related languages, especially in the more abstract (higher) regions and more domain-specific (lower) regions of the OB.

Extension: the LB item is more specific than the appropriate OB item(s); that is, it straightforwardly subcategorizes some OB item(s). In this case, the OB is grown by including the LB item as a child of the OB item. For example, if the OB were initially constructed from English, its system of honorifics would probably contain only two items, one for *FormalSuperordinate* and the other for *EqualOrSubordinate*; a Japanese LB would cause this system to be fleshed out to include a more elaborate substructure along the lines described in [Bateman 88].

Cross-classification: the LB item represents aspects of more than one OB items. This is the case the new language partitions the phenomenon under consideration in a different way to the previous language(s) studied. Typically, several parallel LB items represent one partitioning of the phenomenon and several OB items represent a different partitioning. In this case, two alternative strategies can be followed. The first is to enter the LB items into the OB as a parallel but distinct taxonomization of the phenomenon, and all their descendants must be added as well, unless items representing the same descendants are already in the OB, in which case these items must be linked up also to the appropriate LB item(s). The second strategy is to create the cross-product of the two sets of items. For example, if the OB partitions *Objs* into *UncountableObj* (mass) and *CountableObj* (count) types, and the new language partitions *Objs* into (say) *TallSkinnyObj* and *OtherObj* types, and neither LB class is a proper subset of either OB one, then four new items must be formed: *Countable-TallSkinnyObj*, *Uncountable-TallSkinnyObj*, *Countable-OtherObj*, and *Uncountable-OtherObj*. Every item subordinate to either item in both LB and OB must then be reclassified with respect to the new cross-product items.

One difficult case arises when the same item is taxonomized in the LB and OB under mutually exclusive items. For example, the domain concept *Sand* may be classified as *UncountableObj* in one and *CountableObj* in the other. In this case, the respective LB and OB items *UncountableObj* and *CountableObj* must be differentiated as, for example, *UncountableObj1* and *CountableObj1* and *UncountableObj2* and *CountableObj2*, the cross-product taken, and any subordinate items reclassified. Though this may become a combinatorially

expensive operation in principle, in practise we find it seldom occurs in our domain; furthermore, the fact that all three our current LBs for English contain fewer than 200 non-domain items (see Section 3), and we do not expect significantly larger LBs to be necessary.

2.4 Functional Grounds for Ontology

A certain amount of leeway exists in the construction of the LB. This leeway should be used wisely, since, as mentioned above, similarity with the existing OB can significantly simplify the merging process. For best results, linguistic phenomena should be modeled according to their underlying functional reasons. Functional distinctions often have different syntactic expressions in different languages; it is therefore important to represent in the interlingua the function and not the form (for example, the rule "passive translates to passive" ignores the reason why passive was used in the original and cannot ensure an appropriate target expression). Understanding the true source of variations and creating in the ontology appropriate means for representing them is of central importance.

2.5 Tradeoffs in Ontology Content

The ontology construction method outlined here produces a set of terms explicitly defined to represent all the pertinent linguistic phenomena of each language being covered. For practical purpose, however, this procedure may introduce more complexity than savings in the case where some phenomenon carries meaning only in one of the languages being handled. If for example only one language differentiated *Number* into *Single, Dual*, and *Multiple* (as Arabic and Hebrew do) while all the others just differentiated *Single* from *Multiple*, the *Dual* option can be removed and handled only by the Arabic and Hebrew generators, who in the case of *Multiple* have to determine (somehow) whether the entity in question is dual or not, or whether they should simply generate an alternative locution. For any particular language, the more specifically attuned the LB upper regions are to specific forms of expression, the more syntactically oriented information enters the ontology. As a result, though LB construction for that language is simplified, its merging with the OB is made more complex, as is subsequent merging of the OB with other LBs. In addition, analysis of other languages is complicated, since information must be sought that may not be present in the source text. To top it all, this information is totally irrelevant when translation occurs to languages other than the complexifying culprit.

Easily recognized during step 2(b) of the merging process of LB and OB by the lack of comparable items in the OB, such unique LB phenomena can be omitted from the Interlingua, ignored during analysis, and incorporated during a pre-realization phase in generation, either by using default values, preselected settings, or by turning to a human (in-editor or post-editor) for help. The relative costs of incorporating such phenomena into the LB versus leaving them out and handling them when needed depend on the following:

- the complexity of the phenomenon (which is proportional to the number of LB items required to represent it),
- ease of handling it by default or circumvention,

- the frequency of translation into the language(s) exhibiting the phenomena.

When LB items are swallowed into specific language processors, a record of the removal of the phenomena should be left in the OB at the appropriate point(s) in the taxonomy. Such a record will be encountered during later addition of other languages, enabling arguments to be made for the reintroduction of the phenomenon into the OB itself if appropriate.

In this manner, the Interlingua may be more or less language-neutral, requiring correspondingly more work to generate the more distant ones. Its optimal positioning requires a careful analysis of the tradeoffs involved.

3 Current Work

We recently began constructing an Interlingua for the PANGLOSS machine translation system, using the terminologies developed by the three partners, namely a set of ontologies and domain models developed at the Center for Machine Translation at CMU using the ONTOS acquisition interface [Monarch 89], IR, the Intermediate Representation terminology used at the Computing Research Laboratory of New Mexico State University [Farwell 90], and the Upper Model developed for the Penman language generator at USC/ISI [Bateman et al. 89]. Both CMT's TAMERLAN and IR have already been used to support Interlingual machine translation of several languages, while variants of the Upper Model suited for German, Japanese, and Chinese are under construction at GMD/IPSI (Germany) and the University of Sydney (Australia).

As expected, except for names, we found little disagreement among the three ontologies in their upper regions (while the IR is not explicitly taxonomized into an ontology, this can be done without problem). The most recent ontology from CMU contains approximately 185 items, IR approximately 150, and the Penman Upper Model approximately 190 (the latter two contain only linguistic generalizations, no domain entities).

4 Conclusion

In this paper we addressed the problem of constructing in a principled way one of the three components of an Interlingua for machine translation: the ontology of terms which are used in the Interlingua notation to capture the meaning of the source text and govern the generation of the target text. The methodology described is based not on intuitive grounds about what is and is not 'true' about the world, which is a matter for philosophers and concerns the question of language-independence, but is based instead on more practical concerns, namely what information the analysis and generation programs require in order to perform their tasks.

Given our belief that the a true language-neutral ontology of terms can only be approached asymptotically, the method we outline for constructing the ontology involves a stepwise folding in of one language at a time. Any "syntactic pollution" present in an LB

will either be merged into the ontology and stand explicitly as something other language analyzers and generators have to handle or it will be swallowed in the analyzer and generator for the specific language as a clearly identified item that requires special treatment. In either case, an explicit and declarative representation of the distinctions required to control the analysis and generation of the languages handled results.

Though (depending on the nature of the analysis and generation procedures) a taxonomy of this kind need not always resemble what most people intuitively think of when they talk about an interlingua ontology, we claim this is a moot point, since what we and they are after is a practical construct that can be used effectively in an MT system, and the method of construction outlined in this paper is a way of achieving one.

References

[Bateman 88] Bateman, J.A. 1988. Aspects of Clause Politeness in Japanese: An Extended Inquiry Semantics Treatment. In *Proceedings of the 26th Annual Conference of the ACL*, Buffalo (147-154).

[Bateman et al. 89] Bateman, J.A., Kasper, R.T., Moore, J.D. and Whitney, R.A. 1989. A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Unpublished research report, USC/Information Sciences Institute, Marina del Rey.

[Brugman 83] Brugman, C. 1983. The Use of Bady-Part Terms as Locatives in Chalcatongo Mixtec. In Report no. 4 of the Survey of California and Other Indian Languages, University of California at Berkeley (235-290).

[Carlson & Nirenburg 90] Carlson, L. and Nirenburg, S. 1990. World Modeling for NLP. Technical Report no. CMU-CMT-90-121, Carnegie Mellon University, Pittsburgh.

[Dahlgren et al. 90] Dahlgren, K., McDowell, J., and Stabler, E.P. 1990. Knowledge Representation for Commonsense Reasoning with Text. *Computational Linguistics* 15 (149-170).

[Farwell 90] Farwell, D. 1990. Description of the Intermedia Representation System for the CRL Multilingual Machine Translation System. Unpublished document, Computing Research Laboratory, New Mexico State University, Las Cruces.

[Hinrichs et al. 87] Hinrichs, E.W., Ayuso, D.M., Scha, R. 1987. The Syntax and Semantics of the JANUS Semantic Interpretation Language. In *Research and Development in Natural Language Understanding as Part of the Strategic Computing Program*. Annual Technical Report no. 6552, BBN Laboratories, Cambridge.

[Jackendoff 85] Jackendoff, R. 1985. *Semantics and Cognition*. Cambridge: MIT Press.

[Kay et al. 91] Kay, M., Gawron, J.M., and Norvig, P. 1991. Verbmobil: A Translation System for Face-to-Face Dialog. CSLI Study.

[Lakoff 87] Lakoff, G. 1987. *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press.

[Lang 91] Lang, E. 1991. The LILOG Ontology from a Linguistic Point of View. In O. Herzog and C-R. Rollinger (eds), *Text Understanding in* LILOG. Berlin: Springer (464-480).

[McCawley 78] McCawley, J.D. 1978. Conversation Implicature and the Lexicon. In P. Cole, ed. *Syntax and Semantics* vol. 9. New York: Academic Press (245-259).

[Meyer et al. 90] Meyer, I., Onyshkevych, B., and Carlson, L. 1990. Lexicographic Principles and Design for Knowledge-Based Machine Translation. Technical Report no. CMU-CMT-90-118. Carnegie Mellon University, Pittsburgh.

[Monarch 89] Monarch, I. 1989. ONTOS Reference Manual. Technical Memo, Center for Machine Translation. Carnegie Mellon University, Pittsburgh.

[Nirenburg & DeFrise 92] Nirenburg, S. and Defrise, C. 1992. Application-Oriented Computational Semantics. In R. Johnson and M. Rosner (eds.), *Computational Linguistics and Formal Semantics*. Cambridge: Cambridge University Press.

[[Rosch 73] Rosch, E. 1973. Natural Categories. *Cognitive Psychology* 4 (328-350).

[Whorf 56] Whorf, B.L. 1956. *Language, Thought, and Reality: Selected writings of Benjamin Lee Whorf*, ed. John B. Carroll, Cambridge: MIT Press.