[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

THE EUROTRA (UK) MONOLINGUAL DICTIONARY**

Kerry Maxwell & Blaise Nkwenti-Azeh

Centre for Computational Linguistics
UMIST, Manchester
September 1990

ABSTRACT

This paper provides an overview of the Eurotra-UK Monolingual Dictionary (EMD) - the English lexicon designed and implemented by the British Eurotra group. An outline of some of the design considerations will be given, along with details of the size and scope of the lexicon. A brief description of the feature specifications of entries will also be provided for the major grammatical categories, and the interaction between the dictionary and the grammars will be sketched. Lastly, we will suggest some key areas for research, by looking at problems that need to be resolved in order to reduce redundancy and improve the theoretical soundness of the dictionary.

1. INTRODUCTION

The advent of computers and their subsequent introduction in the field of lexicography has led to renewed intense activity in this domain, and to a re-appraisal of the nature and function of dictionaries. It is even being suggested in certain circles that automation in lexicography has given birth to a new discipline - computational lexicography. The most significant contribution of automation to lexicography has so far been to enhance the versatility of its products, notably, diversifying or varying the structure, layout and contents of printed-page dictionary articles, changing the nature of dictionary consultation, and allowing access to very large information bases. *[Bennett et al., 1986:36f]* Prior to computerisation, the definition of what constituted a dictionary was more or less straightforward (at least from the point of view of content), namely, "an alphabetically ordered list of words along with an explication of their meanings". Dictionaries, glossaries, lexicons were descriptive labels for different lexicographic products. Nowadays, the term 'dictionary' has acquired a much wider extension — sometimes encompassing all the three products above, and much more --; few linguists/lexicographers indeed would give it a straightforward, unqualified definition -- be that in terms of format, content or user-group. It is thus important in any contemporary discussion on computer-held dictionaries or description of a dictionary tool that we begin with a clarification or re-definition of the notion.

2. RE-DEFINITION OF DICTIONARY VIEWS

Existing definitions of 'dictionary' may be subcategorised under two broad divisions: the conventional, "non-computationally-relevant" definition, and the "computationally-relevant" definitions.

** *This paper was presented at a Machine Translation Workshop organised by the Speech and Language Technology (SALT) Club, in UMIST-Manchester, 2-3 July 1990. The research reported on here was undertaken in the framework of the EUROTRA Machine Translation Project, co-sponsored by the UK Department of Trade and Industry/Information Engineering Directorate and the Commission of the European Communities. The views of the authors of this paper are not necessarily those of the EUROTRA Project Management.*

2.1. Non-Computationally-Relevant Dictionary

In the traditional definition of 'dictionary' the focus is, understandably, on structure/format and content, rather than on use. The traditional/conventional view of the 'dictionary' is that it is

> *"A book, usually arranged in alphabetical order and dealing with the words of a language, so as to set forth their orthography, pronunciation, signification, and use, their synonyms, derivation, and history, or at least some of these"* (part-definition of "dictionary" in Shorter Oxford English Dictionary, [*Onions, 1973*])

Alternatively expressed, a dictionary is

> *"A book in which the words of a language are listed alphabetically and their meanings are explained"* (part-definition of "dictionary" in Cobuild Dictionary, *[Sinclair, 1987])*

This definition is still generally valid (at least for the vast majority of printed-page[1] products), and still the most common intension of the notion.

2.2. Computationally-Relevant Dictionary

In a computational environment, we are dealing primarily with so-called "lexical data bases" the reference is much broader, especially as 'word signification' no longer appears to be an essential or intrinsic attribute. The term 'dictionary' is here best defined by use-environment rather than content/format. One can isolate from the vast literature on the subject, the following major — albeit partly overlapping — classes/types of lexical databases:

Type-1: Databases for NLP and MT (existing in machine-form only). These are single-system, application-specific dictionary/lexicon modules; there are as many of these as there are applications or systems.

Type-2: Databases for "monolithic", i.e. single-product, printed-page dictionary production. These are dictionary-entry files which are intended solely for computer typesetting, and which contain output and data management information but little or no data-manipulation information.

Type-3: Databases for diversified-product, printed-page dictionary generation. Databases which fall into this category are advanced versions of Type-2; they have incorporated various data-manipulation techniques/algorithms which allow selective retrieval of information. They have been referred to as dictionary databases (DDB). Two well-known examples are LDOCE [*Procter, 1978*] and the Van Dale Dutch defining dictionary [*Van Dale & Kruyskamp, 1976*].

Type-4: Multifunctional, multipurpose NLP databases. This is the area in which most of the computational lexicography research effort is currently concentrated. Some noteworthy examples of this class of databases include the Italian Machine Dictionary (DMI) conceived from other machine dictionaries which exist as components in MT and IR systems [*Calzolari, 1988*], BONNLEX which is a cumulative word database compiled from 12 other databases [*Lenders, 1986*], the CODER lexicon based on CDEL and W7 magnetic tapes [*Fox et al., 1988*].

---

[1] Until recently the terms "hand-held" and "printed-page" were synonymous. With the introduction of electronic, pocket-size word-banks, it would seem appropriate to differentiate between these two products.

Type-5: Dustbin category which encompasses (spelling/style checkers, morphological analysers, synonym finders, etc.), mainly tools developed for use in word-processing and document-preparation environments.

2.3. Categorising the EMD

The EMD exists primarily as the lexical component/module of an MT system. The dictionary shares some common features with other MT dictionaries; it also contains features that are specifically relevant and unique to the application it supports. In order to place the ensuing EMD description in the proper perspective, it is useful at this stage to relate the dictionary to, and establish its location within the above classification.

For obvious reasons, the EMD is not a Type-5 database: it is neither a word-processing adjunct nor document-preparation add-on. The EMD is equally neither a Type-2 nor a Type-3 database; again, the reasons are not difficult to discern. Firstly, the dictionary does not belong to the same subclass of machine-held, printed dictionary resources as Longmans, Collins, Van Dale, Harraps etc.; secondly, definitions, usage notes, stylistic information, etymological information, synonyms and related types of information commonly found in printed-page dictionaries are either not encoded or not used in any principled way by the system grammars which the EMD supports.

As we have earlier mentioned, the EMD is theoretically and functionally a Type-1 DBD (Database Dictionary). It is, to all intents and purposes, the lexicon module of an MT system, similar say, to the lexicon in such MT systems as the operational TAUM-METEO (Canada), the commercial METAL (FRG), Logos (USA), Atlas (Japan), etc. But, in difference to these or other MT/NLP systems, and as our description of some information categories in the dictionary shows, only a subset of the information contained in the EMD is actually used for the project. The existence of application-independent lexical information offers the EMD considerable scope for exploitation in a different CL environment. In its current state, we view the EMD as a half-way house between the Type-1 mono-functional, single-purpose database and Type-4 multi-functional, multi-purpose database.

3. EMD DESCRIPTION [2]

3.1. Design Considerations

The following design considerations underlie the Eurotra (UK) monolingual dictionary:

(a) MT dictionaries require highly formalized information, as opposed to non-formalized information such as headword and definition, glosses etc. Entries in the EMD are therefore "flat" sets of attribute-value pairs, which can be directly accessed by the grammar modules.

(b) Eurotra dictionaries function under a multilingual environment and are therefore subject to centralized standards of representation. This leads to the notion of "legislated features" (ie: those specified by central legislation as requisites for a language-neutral representation of IS (Interface Structure)), and "non-legislated / language-specific features" (ie: those features which are specified by individual language groups as part of their own lexicon design.)

(c) The stratificational design of the Eurotra framework requires that dictionaries are available at all levels of representation (but cf. 3.4.2.).

---

[2] The following sections assume a basic understanding of the stratificational approach adopted in Eurotra. (cf. [*Durand el al., 1989*]

3.2. Facts and Figures

3.2.1. Vocabulary

The dictionary covers general language vocabulary of all major and minor classes, and terminological vocabulary in the subject field of satellite communications. The current source for our vocabulary is the centrally approved TELECOM corpus, which contains 50 pages of telecommunications texts.

3.2.2. Size[3]

We have a dictionary of around 540 lexical items which is fully integrated with the grammar modules. This constitutes 699 lexical rules which are appended to the grammar at ECS (Eurotra Configurational Structure). Our external lexical resource consists of around 5,000 lexical items. This constitutes something in the region of 9,000 individual entries, of which around 1,500 are terminological.

3.2.3. Storage Medium

After encountering problems with the UNIFY database system earlier in the project, the dictionary is now written and stored in ASCII text files, which are manipulated by simple text-processing facilities.

It is our feeling that the lexicographic workbench which is currently official to the project (hereafter EDB) does not offer any particular facilities which would be attractive enough to renounce the freedom of writing entries in ASCII text files. However, it is our intention to review the use of EDB in the future in terms of its capacity as a structured storage device for our lexicon.

3.3. Some Comparisons

Commercial MT Dictionaries

- LOGOS (USA): 1,000,000 entries
- MacroCAT/MicroCAT (Weidner) : 8,000 stems
- SMART (USA) : 10,000 entries
- ATLAS-1 (JAPAN) : 'Basic' Dict, of 53,000 entries
  'Technical' Dict, of 250,000 entries

Research-based MT Dictionaries

- ENGSPAN (USA): 50,000 words
- SPANAM (USA): 64,000 words
- TAUM-AVIATION : 4000 words
- TITUS (FRANCE): 12,000 words (basic dict. of ,1000)
- XTRA (USA): 7000 headwords

One can see from the figures that manufacturers of commercially-based MT systems have invested in the production of large scale lexicons in order to make their product commercially more attractive.

[3] NB: The figures quoted here represent the official size of our dictionary as delivered centrally to the project. The "in-progress" size of our dictionary is in the region of 17,000 entries (as at *30/09/90).*

In research-based MT (and indeed in NLP in general), it is true that less attention has been paid to the magnitude of the lexicon. During a workshop held in Manchester on linguistic theory and computer applications it was suggested that the average number of entries per lexicon in the NLP systems represented by the participants was only 25, if one of the systems with a larger lexicon was ignored. [*Whitelock et al., 1987:234*]

In the light of such statements the complexity of entries needs to be taken into account. It might be a fair assumption that those systems with reputedly "small" dictionaries may in fact boast lexical entries of particular complexity. Conversely, those systems which claim bigger lexicons may have entries which are merely a coding of lexical unit and category. Indeed, notions such as "headword", "word", "entry" and so on are vague and potentially quite misleading, in that none of them for instance rule out the possibility of referring to "full-forms". Equally, figures quoted for bilingual dictionaries may refer to a simple source and target language lexical mapping.

We should therefore clarify what is meant by the term "entry" in the Eurotra UK Monolingual Dictionary. In fact, because our formalism does not currently provide us with any mechanism for representing alternations, each distinct syntactic realization generally constitutes a separate entry.[4] Therefore, a verb of a particular reading will have entries corresponding to each syntactic frame it is assigned.

The notion of entry clarified, we can nonetheless state that all dictionary entries are fully coded according to our feature theory. This means that entries for major categories would contain at least 10 feature specifications, which are not so highly formalized as to be inaccessible to anyone unfamiliar with the Eurotra formalism.

3.4. Rationale of the EMD

There are two key methodological principles unique to the UK group implementation of the monolingual dictionary: comprehensiveness and level-independence.

3.4.1. Comprehensiveness

The position adopted by the Eurotra-UK group has continually been that the lexicon's feature set will include ALL information deemed necessary to comprehensively describe a lexical item. In other words, the lexicon design has not been restricted to project-specific requirements or the state of the system at any one point in its development. It has rather aimed to include information which may be accessed by the system at a later stage of development, but which in any case provides the transfer writer with additional information about the proper monolingual description of an item.[5]

3.4.2. Level-independence

As stated previously, the stratificational approach in Eurotra requires that dictionaries are available at all levels of representation. However it has been the Eurotra-UK policy to develop a single, level-independent lexicon (i.e., a lexicon which includes all features relevant to each representational level) from which level-specific entries can be automatically generated. This enhances the reusability of our lexicon, in that all lexical information is collated in base entries which are then independent of the stratificational design.

---

[4] An exception here would be the encoding of nouns, since the optionality of noun arguments has led us to formulate noun frames whose elements can be optional; the assignment of a frame indicates the fact that the noun can occur with all (or any number of) the specified arguments or none of the arguments at all.

[5] A free-lance lexicographer was involved in the early design stages of the EMD, and with her expertise a feature *ten* covering the major classes was developed.

4. EMD ENTRY FEATURES

In this section we give example entries for four categories (noun, verb, adjective, preposition) covered in the dictionary, and comment on some of their attributes.

4.1. Some Examples of Level-independent Entries

Entries consist of "flat" sets of attribute-value pairs. Several lines of free text (e.g. glosses, examples, etc.) may appear after an entry. These lines of text are always preceded by "%%".

NOUN

```
question37={
        gb_lu=question,
        cat=n,
        gb_rno=l,
        morph_source=either,
        nclass=common,
        n_morphol=add_s,
        rsf_human=no,
        rsf_loc=none,
        rsf_coll=no,
        det_use=always_det,
        v_agr=sing,
        plurality=standard_pl,
        subconjform=whether,
        ers_frame=subj_sobj_inf,
        gb_isframe=arg 12,
        term=no
}.
%% the (government's) question whether to.....

emphasis12={
        gb_lu=emphasis,
        cat=n,
        gb_rno=l,
        morph_source=verbal,
        nclass=common,
        plural=emphases,
        rsf human=no,
        rsf_loc=none,
        rsf_coll=no,
        det_use=any,
        v_agr=sing,
        plurality=standard_pl,
        ers_frame=subj_objpp,
        gb_pformarg2=on,
        gb_isframe=arg12,
        term=no
}.
%% eg: the industry's emphasis on digital techniques
```

VERB

```
appear120= {
        gb_lu=appear,
        cat=v,
        gb_rno=2
        verb_type=raising,
        cattype=main,
        v_morphol=s_ed,
        passive=never,
        control=no,
        ers_frame=attrsubj,
        gb_isframe=argl2,
        term=no
}.
%% they appear confident

begin230={
        gb_lu=begin,
        cat=v,
        gb_rno=l,
        verb_type=ergative,
        cattype=main,
        v_morphol=irreg,
        pers3sing=begins,
        pres_part=beginning,
        past=began,
        past_part=begun,
        passive=either,
        control=no,
        ers_frame=sobj_inf,
        gb_isframe=argl2,
        term=no
}.
%% he began to withdraw
```

ADJECTIVE

```
simple42a={
        gb_lu=simple,
        cat=adj,
        gb_rno=l,
        frame=s_inf,
        adj_morphol=add_r,
        more=yes,
        npdiacr=pro,
        gradable=yes,
        adjpos=predicative,
        gb_isframe=arg1,
        term=no
}.
%% the cake was simple to make
```

PREPOSITION

```
about = {gb_lu=about,
cat=p,
gb_rno=l,
ptype=standard,
p frame=np
}.

because = {gb_lu=because,
        cat=p,
        gb_rno=l,
        ptype=subconj,
        p_frame=s_tensed,
        clausetype=cause
}.
```

4.2. Some Notes on Features

4.2.1. NOUNS

entryXX={ This is a label or entry id, which constitutes the rule name in the case of lexical rules appended to the grammar.

gb_lu Holds the value of the lexical unit.

cat = n Grammatical class. In the case where an item potentially belongs to more than one class, we code only that instantiated in the corpus.

gb_rno This is a purely SEMANTIC reading number which is used to group together entries of the same semantic reading, i.e. there may be a set of entries which differ with respect to some syntactic feature but which all have the same reading number. In the example entries for the noun "relationship" below, the first two illustrate syntactic variants (and so do not differ with respect to reading number), whereas the third entry illustrates a different semantic interpretation (and therefore has a different reading number).

```
'relationshipl'={
        gb_lu=relationship,
        cat=n,
        gb_rno=l,
          …
          …
        blgbers_frame=classpp_conj,
        gb_pformargl=between,
        gb_isframe=argl,
        term=no
}.
%% the relationship between X and Y

'relationship2'={
        gb_lu=relationship,
        cat=n,
        gb_rno=l,
          …
          …
        blgbers_frame=classnp_pp,
        gb_pformarg2=to,
        gb_isframe=argl2,
        term=no
}.
%% the relationship of X to Y
```

```
'relationship3'={
        gb_lu=relationship,
        cat=n,
        gb_rno=2,
           …
           …
        blgbers_frame=classpp,
        gb_pformargl=with,
        gb_isframe=arg1,
        term=no
}.
%% john''s relationship with mary
```

morph_source Denotes whether the noun has a relationship with a verb, an adjective, both, or neither. NB: this does not necessarily mean that the noun is derived from the verb/adj. However it proves useful in conjunction with frame assignment; eg: if morph_source is verbal then it may be that clues about the appropriate frame assignment can be gleaned by looking at the related verb.

eg: development - develop

The government's development of proposals
The government developed proposals

congruence - congruent

the congruence of A with B
A is congruent with B

nclass This is a subclassification of nouns into 'common', 'proper' , 'classifier' (eg: fractions and percentages) 'date' and 'measure' (eg: kilometres etc)

n_morpho1 This is a morphology feature which specifies values for the formation of plural forms of nouns. (This value is now used by the morphology tool.)

plural This feature holds the plural form of a noun explicitly where this is irregular, eg: emphasis - emphases

We have a set of restricted semantic features ('rsf's') for nouns as follows, which are not yet implemented within the system (but: rsf_loc=time is used in the analysis of temporal adverbs.)

rsf_human The noun refers (or does not refer) to humans/people

rsf_loc Values are 'space' (concrete), 'time', 'unmarked' (for many deverbals), 'none' (abstract)

rsf_coll The noun refers (or does not refer) to a collective body (eg: "committee", "organization")

det_use	Specifies the occurrence of determiners with the (singular) noun; eg: "the_nodet" means this noun can take "the" or occur without a determiner but cannot take "a" - eg: "personnel"; "always_det" means always takes a determiner, this can be "a" or "the" eg: "satellite" .
v_agr	Specifies agreement with verb:- eg: 'sing' - takes singular verb in base form. 'plu' - takes plural verb in base form eg: "trousers", 'either' - takes either eg: "government"
plurality	Gives indications as to whether it is possible to form the plural of a noun, and if so, whether this is formed in a "standard" way. To clarify, example values are "no_pl" (i.e. it is not possible to pluralise this noun, e.g. "personnel") or "standard_pl" (i.e. this noun pluralises in a standard way, whether as an irregular form or using a regular ending).
gb_compform	Specifies complementizers where appropriate; eg: "belief that", "question whether".
ers_frame	Specifies syntactic frame of the noun. NB: We have optionality within noun frames so that the assignment of a frame indicates the fact that the noun can occur with all (or any number of) the specified arguments or none of the arguments at all.
gb_pformargX	Where X is the number of the argument. Specifies what are considered argument bound prepositions, eg: "emphasis on ...", "compliance with"
gb_isframe	Specifies arity. (direct mapping with ers_frame)
term	term "flag" (ie: 'yes' or 'no')

4.3. VERBS

cat = v

cattype	This is " main" - for any verb which is not considered a modal, or "modal".
v_morphol	Specifies value representing inflectional pattern where this is regular, or marks irregular verbs.
pers3sing	This feature, and the following three, hold corresponding irregular forms, where appropriate.
prespart	
past	
pastpart	
verb_type	Values are 'reflexive', 'raising' (eg: "seem"), 'ergative' (eg: "accelerate") ,'report' (with direct speech eg:- "demand"), 'dummy_obj' (eg:- "I find it amazing"). A combination of these may be given.

passive — Specifies whether the verb can occur in the passive (ie: "either" [active or passive] or "never") or whether it always occurs in the passive (ie: "always").

control — If the verb is a control verb, specifies whether subject (eg: "promise"), object (eg: "persuade"). Value "no" if not a control verb.

ers_frame — Specifies syntactic frame assignment.

gb_compform — Specification of complementizers where appropriate, eg:- "believe that", "wonder whether".

gb_isframe — Arity, direct mapping with ers_frame.

gb_pformargX — Specification of argument bound prepositions, eg:- "coincide with", "concentrate on."

term — (as for nouns, see above).

4.4. ADJECTIVES

cat = adj

adj_morpho1 — Specifies values which indicate the formation of the comparative and the superlative where this is regular.

compar — Holds the value of the comparative where this is irregular.

super1 — Holds the value of the superlative where this is irregular.

more — Specifies whether the comparative and superlative can be made by using "more" and "most".

npdiacr — Specifies whether the adjective can stand as a proform for a noun, eg: "the poor"

gradable — Specifies whether the adjective can be graded eg: "very beautiful", "quite interesting"

adjpos — Specifies positional information:-

eg:- 'premod' - premodifies a noun eg: "main"
'predicative' - only appears in predicative position
'either' - premodifying or predicative
'postnoun' - only appears after a noun eg: "president elect"

adjclass — Could be considered a syntactico-semantic feature in that it specifies information about the scope of the adjective but is intended for use in generation to determine surface syntactic positions.

eg:- 'qual' - qualitative adjs. (occur before colour and class)
'class' - classifying adjs. (occur closest to noun which they modify, ie: after 'qual' and 'colour')
'colour' - colour adjs. (occur after 'qual' but before 'class')

eg:- "huge black medieval castle"

adj_type — Used to specify special behaviour of adjectives

eg:- adj_it - adj. occurs in construction with "it" as dummy subject, "it is doubtful whether...."

recip - adj. behaves in a similar way to recip. verbs
eg:- "x is equal to y", "x and y are equal"

standard - default value, neither of the above.

adj_control — Used to specify whether adj. implies subj. or obj. control.

eg:- subj_is_subj "John is eager to please"
subj_is_obj "John is easy to please"
no not control...

discon — Specifies whether the adjective takes discontinuous arguments, eg: "interesting", as in "this book is interesting to read" also has a discontinuous realization: "an interesting book to read".

gb_compform — Specifies complementizer, where appropriate, eg: "doubtful whether", "strange that".

frame — Specifies syntactic frame; note that this is merely a surface syntactic analysis of the adjective's argument structure, as opposed to ers_frames for nouns and verbs, which attempt to express deeper syntactic relations.

some brief examples:-

s_inf "difficult to see"
pp_s_inf "simple for him to correct"
pp1_pp2 "dependent on Jim for help"
pp_s_tensed "it is apparent to me that this is wrong"
s_ing "I'm happy doing this"

gb_pformargX — Specification of what are considered to be argument bound prepositions, eg: "dependent on", "particular to", "aware of.

gb_isframe — Specifies arity, direct mapping to ers_frame.

term — (as above, see Nouns)

4.5. PREPOSITIONS

cat = p

ptype — In the framework, some subordinating conjunctions are regarded as prepositions. To mark this, ptype can be:- 'standard' or 'subconj'.

p_frame — Specifies a surface syntactic analysis of prepositional arguments, eg:-

np - "about satellites .."

pp - "from above the window ..."

clausetype — If ptype is is 'subconj' then clausetype is also specified. Values are 'concess' (which means that concessive clauses are introduced by this preposition, eg: "although"), 'cond' (to indicate that conditional clauses are introduced by the preposition, eg: "unless"), and 'cause' (for causative clauses introduced by the preposition eg: "because").

5. INTERACTION OF LEXICON AND GRAMMAR MODULES

5.1. ECS

Until quite recently, a conversion routine took level-independent entries, and, on the basis of the morphology feature assigned, generated a set of full-form entries enriched with appropriate surface features, which were then appended to the grammar as lexical rules. Some examples follow. Note the appearance of 'gb_vform' and 'finform', which are surface features for verbs.

appear118_l={gb_lu=appear,cat=v,gb_rno=l,cattype=main,ers_frame=subj,
term=no,lex=appear,gb_vform=infin}.

appear118_2={gb_lu=appear,cat=v,gb_rno=l,cattype=main,ers_frame=subj,
term=no,lex=appear,gb_vform=finite,finform=pres}.

appear118_3= {gb_lu=appear,cat=v,gb_rno=l ,cattype=main,ers_frame=subj,
term=no,lex=appears,gb_vform=finite,finform=tsg}.

appear118_4={gb_lu=appear,cat=v,gb_rno=l,cattype=main,ers_frame=subj,
term=no,lex=appearing,gb_vform=prespart}.

appear118_5={gb_lu=appear,cat=v,gb_rno=l,cattype=main,ers_frame=subj,
term=no,lex=appeared,gb_vform=finite,finform=past}.

appear118_6= {gb_lu=appear,cat=v,gb_rno=l,cattype=main,ers_frame=subj,
term=no,lex=appeared,gb_vform=pastpart}.

This amounted to 1,704 ECS lexical rules (l-rules). The approach was relatively easy to implement, and produced an effective runnable lexicon, but was disadvantageous in that it led to longer loading and compilation times for the grammars. It also made the addition of new lexical items to the 1-rule set rather cumbersome, since these were usually done without the aid of the conversion routine, which could only be usefully implemented batchwise.

However, we have recently implemented a morphological analyser, which dispenses with the need for conversion to full-form entries. Level-independent entries are incorporated directly into the ECS grammar, and values assigned to the morphology feature invoke a set of morphological rules. These generate a virtual set of full-form entries which are active as a surface string is parsed. The current number of 1-rules is therefore 699. Grammar compilation and loading times are significantly improved, and making extensions to the runnable lexicon is a straightforward importation of level-independent entries.

5.2. IS

IS entries are basically the same as level-independent entries, although they do not include morphological information. The "non-legislated" language-specific features (which appear in addition to those necessary for the analysis implementation itself) are usually "commented out", so that the entries compile against the IS feature definition file, but remain available for the information of transfer writers. An example follows:-

```
'emphasisl2'={
        gb_lu=emphasis,
        cat=n,
        gb_rno=l,
        %%morph_source=verbal,
        nclass=common,
        rsf_human=no,
        rsf_loc=none,
        rsf_coll=no,
        %%det_use=any,
        %%v_agr=sing,
        %%plurality=standard_pl,
        ers_frame=subj_objpp,
        gb_pformarg2=on,
        gb_isframe=argl2,
        term=no
 }
```

6. LEXICON AND GRAMMAR

The analysis and synthesis grammars are continually modified as the linguistic structures handled by the system are expanded. The dynamic state of the grammar may have negative consequences for the lexicon. As the grammar is extended it may require additional information from the lexicon, and if this is not already available, non-trivial amendments to the lexicon design need to be made.

In our design and implementation of the EMD we have managed to resist this problem by providing a feature set which contains more than the minimum information needed at any one point in the system's development. In the more advanced stages of the grammar's development, it has drawn upon information which already exists in the lexicon. An example would be the use of the feature "adjpos" (which specifies surface syntactic position of adjectives, whether predicative, attributive etc.) in connection with adjective complementation; viz. a necessary condition for the occurrence of the frame "pp" for adjectives (NB: "pp" meaning that this adjective selects a prepositional phrase argument) was that the value for "adjpos" was "predicative".

One could therefore claim in fact that the more comprehensive the feature set of the lexicon from the outset, the less susceptible the lexicon is to changes invoked by the demands of the grammar. In other words, the grammar can expand to touch upon all the information in the lexicon, rather than the lexicon being forced to expand in order to meet the requirements of the grammar.

7. PROBLEMS AND PERSPECTIVES

The Eurotra lexicographer, it has been remarked, works very differently from the 'classical' lexicographer. Traditionally, the lexicographer examines text corpora to analyse, classify and describe the various meanings of the word. In Eurotra, the lexical entry should primarily contain those syntactic features that are needed by the grammar to analyse and generate syntactically correct sentences. Other aspects are under-developed.[6]

In the remaining sections of this paper we discuss some of the dictionary-coding problems in the project, and outline some of the so-called "under-developed" aspects. We however leave unanswered the

[6] These remarks were made by the working group on "Lexicon and LDB", during a recent Eurotra Workshop held at Noordwijkerhout-Holland, 28 May - 2nd June, 1990.

question as to what extent the deficiencies/limitations of translation output could be attributed to this under-development.

7.1. Coding Efficiency (Redundancy)

A major representational problem faced in the EMD is, not surprisingly, that of REDUNDANCY. The corpora on which the EMD entries are based - Satellite Communications handbooks - were written by different specialists, and in an unconstrained language; one consequence of this is the diversity of style and terminology encountered in the merged corpus. (Note that stylistic and terminological inconsistency are not mutually exclusive.) At the EMD level, one of the manifestations of stylistic inconsistency is the way compounds, abbreviations, and acronyms are written. Variations in writing style can lead to undesirable duplication in a word-based dictionary database.[7]

7.1.1. Compounds

Compound words (including compound terms) in English are written either as a single word (e.g. *earthstation, beamwidth, downlink),* hyphenated *(earth-station, beam-width, down-link),* or as separate words *(earth station, beam width, down link).*[8] The terms "solid", "hyphenated" and "open" respectively are sometimes used to differentiate the three patterns [*Ilson, 1988:76*]. We may point out that slashes are also often encountered (e.g. *carrier/noise ratio, Earth/space link*).

7.12. Abbreviations and Acronyms

As far as abbreviations are concerned, these are written with inter-letter stops, i.e. dots (e.g. *v.h.f., e.i.r.p.),* without stops *(vsli, DC),* or with slashes *(C/N* as in *C/N ratio, AC/DC).* There is also letter-case variation with the alternate use of lowercase and uppercase characters (e.g. *FSS, fss).* Acronyms, in particular, occur in three orthographic forms: uppercase, capital-initial, and lowercase (e.g. *INTEL-SAT, Intelsat, intelsat).* The problem is made worse by the fact that various combinations of the above are encountered in long compounds (e.g. *ground-to-air_t.d.m., hf._radio, half_offset_qpsk, ots satellite, e_&_m_lead_(signalling), e_and_m lead_signalling).*

It is evident from the examples that there can be (indeed, there currently is) a significant amount of redundancy in the dictionary where orthographic variants are all coded as separate entries.

The hyphen vs interword-space variation in compounds is currently resolved by adopting a standard convention in the coding (in the EMD, all hyphens and IW-spaces are replaced with an underscore).

The single-word vs multi-word variation, on the other hand, is much more intractable and is as yet unresolved. The result is that where single-word and multi-word variants of a compound are encountered in the corpus, e.g. "earthstation" and "earth_station", both forms are entered in the dictionary, with identical feature-value information except for gb_lu=.

This means, for example, that "earthstation" and "earth_station" constitute separate entries. The same approach is adopted for variant cases and punctuation, where typographical variants of the same word/term are separately encoded.

Two summing-up remarks on coding efficiency and the size of the lexicon:

[7] In the examples, we ignore s/z spelling variations.

[8] The irrationality in the choice of convention is revealed by the following frequency figures in our corpus:

beamwidth (f=340)	downlink (f=6)
beam width (f=6)	down-link (f=225)
beam-width (f=l)	down link (f=110)

(1) So long as we are dealing with a small, constrained lexicon (and/or users are aware of the dictionary constraints), it is possible to adopt certain ad hoc solutions to the problems created by inconsistency. If, on the other hand, we are dealing with a large lexicon and/or open-ended system, we cannot afford to adopt ad hoc solutions to these problems. It is then necessary to incorporate a text regularisation program, so that the system resolves any orthographic variants or inconsistencies before it processes a user's request.

(2) The current situation with the dictionary is that it feeds directly into the grammars. As previously observed, changes in the grammar (which in fact are very common) can sometimes lead to modifications in the dictionary. So long as the dictionary remains small, this does not pose a serious problem. However, frequent changes to a large lexicon can be costly in more ways than one. A multi-functional multi-purpose DBD has to be shielded from the vagaries of individual applications; the need for a mapping program of some sort or a "dictionary server" [*Kay, 1984:461*] will eventually become apparent, once the results of ongoing research are fed into the dictionary.

7.2. Under-Developed Aspects of the EMD

Two important areas that have not been sufficiently addressed in the EMD are a) the representation of semantic-type information and b) representation of terms.

7.2.1. Semantic-type information

It is now commonly acknowledged by researchers in the MT community that lexical semantic features can and would play an important role in, among other things, disambiguating lexical items, determining lexical selection for transfer, controlling prepositional (or other) attachment, etc. [*Durand el al., 1989*]. Various monolingual LSF (as opposed to "euroversal", i.e. one shared by all language analysis and generation components) schemes are currently being explored [*Zelinsky-Wibbelt, 1986; Togeby, 1988*].

7.2.2. Terms

The problem of representing terms is not specific to the EMD, but rather reflects the approach adopted in the Eurotra project as a whole. A 'term' or 'terminological unit' is, strictly speaking,

> *"any linguistic sign or lexical unit which, within the domain of special languages, has a special and ideally uniquely definable reference, standing in at least one conceptual relationship to another term, and about which assertions can be made and inferences drawn, on the basis of analyses of its constituent elements, its characteristics/ properties, or the relations it contracts with other terminological units of the system."* [*Nkwenti-Azeh, 1989:51*]

The Eurotra approach to terms is founded on at least three common fundamental misconceptions:

(a) Terms are language universal (or in the context of Eurotra, "euroversal"). The difficulty experienced by other language groups in finding interlingual equivalents of English terms underlines this fallacy.

(b) Terms are univocal in reference, in other words, there is a one-to-one mapping between concept and linguistic form. On the contrary, there is evidence of a large degree of contextual synonymy resulting from term-reduction or shortening. In our corpus, for example, the following are used interchangeably:

e_&_m_lead_(signalling)
e_&_m_signalling
e_and_m_lead_signalling
e_and_m_signalling

Term-reduction is also encountered in other special subjects, as seen in the following examples taken from the field of automotive engineering:

connecting rod small end bush
small end bush
bush

exhaust valve lifter cable
exhaust lifter cable
cable

Although it would appear that the problem of term-reduction is analogous to that of say, pronominal anaphoric reference, it is, in reality, far more complex and its resolution has to be approached differently.

(c) A third misconception (arguably unique to Eurotra) is that terms, the majority of which are compounds, behave in exactly the same way as, and therefore can be given the same treatment as general language compounds. Fortunately, opinions are changing about the latter view of terms. Ongoing experimental work within Eurotra does not treat terms as just another set of compounds which happen to occur in only a restricted domain. Nevertheless, it should be said that because of the relative infancy of sublanguage research, the nature, behaviour and use of terms is as yet not sufficiently investigated and still not understood by the majority of dictionary writers; consequently, terms are not given the separate examination they require.

Our current thinking is that more research effort has to be directed at investigating, on the one hand, how terms behave as integral linguistic units, and on the other hand, how terms, as ordinary language items, can be dealt with in a computational environment.

8. CONCLUSION

The future of the EMD is quite promising since we are continually expanding the size of the dictionary, developing better techniques for storage and access of our lexical data, and attempting to enhance the feature set. Although some groups of researchers are now investigating the importation or adaptation of information from existing DDBs (in particular, the LDOCE and OALD databases) [cf. ongoing research by *Boguraev & Briscoe* (LDOCE), *Akkerman et al.* (LDOCE & OALD), *Fontenelle et al.* (LDOCE)] the signs are that considerable effort would be required to make these resources usable in a particular MT system; MT and NLP system developers will continue to develop independent MT dictionaries, unless perhaps, computational linguists and printed-dictionary designers/producers work more closely to facilitate the development of a generic dictionary database which can be easily adapted for NLP/MT applications.

REFERENCES

Akkerman, E., Voogt-van-Zutphen, H. & Meijs, W. (1988) *A Computerized Lexicon for Word Level Tagging.* Amsterdam: Rodopi B.V.

Ananiadou, E., Antona, M., Crookston, I., Juarez, L., Lindop, J., Maxwell, K., Syea, A., Thorpe, S., Underwood, N. & Way, A. (1990) *Eurotra UK Implementation Report, 1 July 1989 - 31 January 1990.*

Bennett, P.A., Johnson, R.L., McNaught, J., Pugh, J., Sager, J.C. & Somers, H.L. (1986) *Multilingual Aspects of Information Technology*. Gower.

Boguraev, B. & Briscoe, T. (1989) Utilising the LDOCE Grammar Codes. In *Computational Lexicography for Natural Language Processing.* Boguraev, B. & Briscoe, T. (eds.), pp.85-116. London & New York: Longman.

Calzolari, N. (1988) The Dictionary and the Thesaurus can be Combined. In *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks.* Evens, M.W. (ed.), pp.75-96. Cambridge: Cambridge University Press.

Durand, J., Bennett, P., Allegranza, V., van Eynde, F., Humphreys, L., Schmidt, P. & Steiner, E. (1989) Linguistics for MT: The Eurotra Linguistic Specifications. To appear in *Machine Translation* (special Eurotra issue).

Fontenelle, T. & Vanandroye, J. (1990). Retrieving Ergative Verbs from a Lexical Database. Forthcoming in *Dictionaries - Journal of the Dictionary Society of North America.* Bailey, R. (ed.).

Fox, E.A., Nutter, T.J., Ahlswede, T. & Evens, M. (1988) Building a Large Thesaurus for Information Retrieval. In *Proceedings of the 2nd Conference on Applied NLP*, February 1988, Austin- Texas, pp.101-108. Association for Computational Linguistics.

Ilson, R. (1988) Contributions to the terminology of lexicography. In *ZuriLEX '86 Proceedings. Papers read at the EURALEX International Congress,* University of Zurich, 9-14 September 1986. Snell-Homby, M. (ed.) (1988), pp.73-80. Tubingen: Francke Verlag.

Japan Electronic Industrial Development Association (JEIDA) (1989) *A Japanese View of Machine Translation in light of the Considerations and Recommendations reported by ALPAC, U.SA.* Tokyo: JEIDA.

Kay, M. (1984) The Dictionary Server. In *Proceedings. 10th International Conference on Computational Linguistics (COLING '84)* 2-6 July 1984, Stanford University, California, p.461. Association for Computational Linguistics.

Lenders, W. (1986) Data Sources for a German Lexical Knowledge Base. Pisa Workshop on *Automating the Lexicon,* 15-23 May, 1986. (Unpublished)

Nkwenti-Azeh, B. (1989) *An investigation into the structure of the terminological information contained in special language dictionaries.* PhD Thesis, University of Manchester.

Onions, C.T. (ed.) (1973) *Shorter Oxford English Dictionary on Historical Principles.* 3rd edition, In 2 volumes. Oxford: Clarendon Press.

Procter, P. (ed.) (1978) *Longman Dictionary of Contemporary English.* London, England: Longman.

Sinclair, J. (editor in chief) (1987) *Collins COBUILD English Language Dictionary.* London & Glasgow: Collins.

Togeby, O. (1988) *A proposal for an extended feature system.* Copenhagen: Eurotra-DK.

Van Dale, J.H. & Kruyskamp, C. (compilers) (1976) *Groot Woordenboek der Nederlandse Taal.* 10th edition. Den Haag: Nijhoff.

Whitelock, P., Wood, M., Somers, H., Johnson, R. & Bennett, P. (1987) *Linguistic Theory and Computer Applications.* Academic Press.

Zelinsky-Wibbelt, C. (1986) An empirically based approach towards a system of semantic features. In *Proceedings of COLING 1986,* Bonn, pp.7-12.