Noun Phrasal Entries in the EDR English Word Dictionary

A. Koizumi, M. Arioka, C. Harada, M. Sugimoto
Japan Electronic Dictionary Research Institute, Ltd.
Tokyo, Japan

L. Guthrie, C. Watts
Computing Research Laboratory
New Mexico State University
Las Cruces, NM, USA

R. Catizone, Y. Wilks
Department of Computer Science
University of Sheffield
Sheffield, UK

Keywords: Lexicon construction, universal features, resources for CL.

1. Introduction

The dictionary construction project at the Japan Electronic Dictionary Research Institute, Ltd. (EDR) in Tokyo began in 1986 and is almost certainly the largest lexicon construction project for computational purposes in the world. This paper describes some aspects of the construction of the English language dictionary, in particular a project to verify and enhance information on noun phrases in the English Word Dictionary undertaken by the Computing Research Laboratory at New Mexico State University and the University of Sheffield. We believe the work so far raises issues of wider linguistic interest which require practical solutions so that the large scale lexicon project can proceed. We hope that this paper will show the complexity, diversity, and richness of the content of the EDR English Word Dictionary.

The key idea has been to construct a system of features, categories and structures for encoding English words and phrases that is, at the same time, universal, or at least sufficiently universal to code both English and Japanese, two very different languages indeed. This is particularly evident in the use of left and right "adjacency attributes" in both the English and Japanese dictionaries. This general idea is a very natural outcome of the general state of linguistic theory, at least in the generative tradition, in its broadest sense: one which emphasises universality in its feature sets and structural constraints, but which has also evolved by a long and tortuous route to the current position where the lexicon is primary in a linguistic system, and all other levels of linguistic analysis can be seen as a projection from that level. The alphabet-soup grammar theories that are now current all share that assumption to some degree.

Thus, a practical attempt to construct a lexicon on principles as universal as possible for computational use, is indeed a project broadly consistent with the state of generative theory. Almost all other lexicon construction projects under way with computation as a main goal (e.g. COMLEX, CUP, Procter 1992 and see Wilks, Slator and Guthrie, in press) are designed principally for English, although CUP intends to augment its structures from non-English corpora as soon as is feasible, and a COMLEX for Spanish is already under discussion. Nonetheless, the sheer scale of the EDR enterprise (see below) and its explicitly universalist assumptions do make it unique. We will now outline briefly the general structure of the dictionaries in the project and then proceed directly to some of the theoretical and computational choices that have been made in the English lexicon.

2. The EDR Dictionaries

The EDR Electronic Dictionary (EDR, 1993; Yokoi, 1990) is designed as the first true machine-readable dictionary that contains, in a readily accessible form, the information required for a computer to understand and generate natural language. As such, the EDR Electronic Dictionary is intended to be universally applicable and is not restricted to a particular application system. The part of the dictionary that handles surface information is kept separate from the section that handles semantic information: surface information that is heavily dependent upon a particular language is stored in the Word Dictionary, and semantic information is stored in the Concept Dictionary.

There are four different dictionaries that comprise the EDR Electronic Dictionary: the Word Dictionary, the Concept Dictionary, the Co-Occurrence Dictionary and the Bilingual Dictionary. The different dictionaries that make up the EDR Electronic Dictionary and the EDR Corpus are set in a structure of mutual interrelatedness. Four types of constituent data are contained in the EDR electronic dictionaries: word entries, concept entries, co-occurrence entries and bilingual entries. Word entries consist of headwords, grammatical information that indicates the grammatical characteristics of the word, and concept identifiers that indicate the concepts represented by a given word in different contexts. Concept entries represent the relationship between two dif-

ferent concepts. Co-occurrence entries use co-occurrence relation labels to describe the possible co-occurrence relations between headwords. Bilingual entries describe the word correspondences between headwords in different languages. Thus, each of the EDR electronic dictionaries is related to the others, and by using the different component dictionaries as a single entity, they can usefully be applied to many forms of natural language processing.

3. The EDR Word Dictionary

The role of the Word Dictionary is to provide morphological, syntactic and some semantic information: the Word Dictionary is divided into a General Vocabulary Dictionary and a Technical Terminology Dictionary and the former is further subdivided into a Japanese General Vocabulary Dictionary and an English General Vocabulary Dictionary, each of which contains 200,000 words. The vocabulary covers words, compounds, and idioms used in ordinary documents. The Technical Terminology Dictionary covers words or terms that are specific to information processing and related fields, and is also split into a Japanese Technical Terminology and an English Technical Terminology Dictionary. Each contains 100,000 words.

The main characteristics of the General Vocabulary Dictionary are:

(1) surface level information and deep (semantic) level information are stored separately;

(2) surface level information is described independent of any specific application system or algorithm;

(3) a large-scale vocabulary contains lexical items used in general writing.

The Word Dictionary is a collection of word entries that contain entry information as shown in Fig. 1.

Fig. 1 Structure and Content of Word Entries

Headword Information	Grammatical Information
Headword Notation	Part of speech
Adjacency Attributes	Syntax tree
Extra Notation	Inflection
Pronunciation	Grammatical attributes
	Function word information
Semantic Information	Supplementary Information
Concept identifier	Usage
Concept illustration	Frequency

The Headword Information provides headword, extra notation, and pronunciation. A headword consists of notation (the orthographic spelling of a word - containing all the characters common to all inflected forms of the word) and adjacency attributes. For phrasal entries, the headword is a list of the pairs of notation and adjacency attributes of each constituent of the phrasal entry. The adjacency attributes indicate the possibility of joining one morpheme to another and are used to create adjacency rules for morphological analysis and generation. EDR employs a bidirectional connection grammar which divides the adjacency constraint attributes into possible connectivity to the left of the word and possible connectivity to the right of the word. This information is not normally described in this form for English, but EDR employs the same method in both Word Dictionaries so that morphological analysis of Japanese and English can be made by the same algorithm. The extra notation information stores headwords in kana for entries in the Japanese Word Dictionary and in a character string form with syllable markers for hyphenation for entries in the English Word Dictionary.

Grammatical information consists of part of speech, syntax tree, inflection information, grammatical attributes and function word information. The grammatical information can be used to find the syntactic structure of a sentence in syntactic analysis. A syntax tree is provided for compound words or idioms consisting of multiple words. The function word code corresponding to the notation of the headword is provided for function words.

A concept (listed under semantic information) in addition to being a fundamental component of the Concept Dictionary, describes the semantic content of any word entry in the Word Dictionary. If the same headword has two or more different concepts, separate word entries are used in the Word Dictionaries. This information is used to distinguish between the various meanings a given word may have. The concept is the link between the Word Dictionary and the Concept Dictionary.

Supplementary information provides information on the usage as well as the frequency of the headword entry.

4. Noun Phrase Entries in the EDR English Word Dictionary

Portions of the English Word Dictionary have been subjected to rigorous verification through projects at the Computing Research Laboratory (CRL) at New Mexico State University, and the University of Sheffield. Following is a report on one phase of the verification project at CRL which was aimed at describing the grammatical information as well as verifying the morphological information. The objects of this phase of the verification project consisted of 37,039 entries initially coded by EDR as noun phrase expressions. Among these entries, 2,389 were treated as single word entries and 34,650 were treated as

phrasal entries (see below for the distinction between single word and phrasal entries).

4.1 Phrasal Entries vs. Single Word Entries in the EDR English Word Dictionary

In the EDR English Word Dictionary, headwords are treated as either single lexical item, while 'phrasal' refers to a word that is composed of more than one lexical item. In addition to the difference made on lexical units, some words are 'treated as' single word entries even though they are composed of more than one lexical item. The type of information that is provided for headwords varies according to the type of headword. Phrasal entries are given the same information given to single word entries but they are also coded with additional information that indicates their internal syntactic structure. The adjacency attributes and the grammatical attributes are given to each of the constituents of the phrasal. Phrasal expressions treated as single word entries are not segmented into constituent words. Included under those words that are treated as single word entries are the following types of words:

-foreign words
-proper nouns
-common nouns derived from proper nouns ("New Mexican")
-idiomatic expressions which do not fit into a generalized phrase structure pattern ("on the cheap," "open sesame")
-function word equivalents

4.2 Information Provided for Noun Phrase Entries

For noun phrase entries in the English Word Dictionary information is provided for the phrase as a whole as well as for the individual constituents that comprise the phrase. Whole phrase information includes designation as either a common noun or proper noun (proper nouns are treated as single word entries and constituents are not separately analyzed), countability, collectivity, gender, verb agreement and article usage. In addition, the head noun is designated.

The constituent information provided for phrasal entries includes left and right adjacency attributes, part of speech, inflection information, and grammatical attributes. The grammatical attributes that are provided for each of the constituents varies according to the part of speech of the constituent. Information regarding collectivity and countability is provided for nouns and information on possible comparative, superlative, or positive degree forms is given for both adjectives and adverbs that appear as constituents of the phrase. Constituent information is provided within the context of the whole phrase. Syntax trees are also described for noun phrase entries.

4.3 Aim of Verification

4.3.1 Coding of Grammatical Information

The primary objective of the verification project was to code the grammatical information for the noun phrase entries. The specific information given for the noun phrase entries included determining the intra-phrasal structure of the phrasal, the grammatical attributes of the constituents and also the grammatical attributes of the entire noun phrase.

4.3.1.1 Syntactic Relationship Between Constituents

The basic principle used in coding intra-phrasal syntactic information is that the information should clarify the syntactic structure of the phrasal entry. For example, the following phrases look similar on the surface, i.e. adjective + noun + noun, but actually the internal syntactic structure of each phrasal is different.

(iii) traveling post office (iv) dead letter box

The adjective "traveling" modifies the noun phrase "post office" in the phrase "traveling post office" while the noun phrase "dead letter", composed of an adjective and a noun, modifies the noun "box" in the phrase "dead letter box."

For building a source of lexical information to be used in language processing, indication of the head noun of a phrase is useful as hypernym information. Location of the head noun cannot be determined automatically from the phrase structure. That is to say, it is often the case that the head noun is the noun occurring in the final position of a phrasal composed of two (or more) lexical items, but this rule does not always apply as there are also cases in which the head noun is the first noun of the phrasal e.g., court martial.

During the actual task, the distinction between the syntactic relationship between constituents was carried out by indicating the intra-phrasal syntax by parenthesizing the immediate constituents with categorical labels. The categorical labels used to mark the grouping of the phrasal are shown in the example below:

EAJ(traveling)/EN1(post)/EN1(office)
-> EAJ(traveling)/EN1(EN1(post)/EN1(@office))

In this syntactic notation a slash (/) divides constituents at the same level, EN1 is an English common noun, EAJ an English adjective, etc.; and the bracketing structure is a linearized tree in a standard form, e.g., in (iii) above the tree expands to the right, while in (iv) it expands to the left. The symbol "@" indicates the head noun.

4.3.1.2 Grammatical Information for the Constituents

Once the intra-phrasal syntax structure of the phrasal has

been determined, the inflection information and grammatical attributes of the constituents are determined. The grammatical attributes of the constituents are determined by considering the constituent as part of the phrase. Given the information of the constituent words coded as separate dictionary entries in the EDR English Word Dictionary, the coding is given based on the behavior of the constituent when it is used in the phrase.

traveling:EAPOS;EANOCMP;EANOSUP
-> EAPOS;EANOCMP;EANOSUP
post: ENSG;ECN1;ENC -> ENSG;ENU
office: ENSG;ECN1;ENC -> ENSG;ECN1;ENC

The coding ECN1 (takes plural ending -s) and ENC (Countable) is changed to ENU (Uncountable) to indicate that the word "post" when used in the context of the phrase, does not inflect.

4.3.1.3 Grammatical Information for the Noun Phrase Unit

The final process in the coding of the syntactic information for noun phrasals involves marking the grammatical attributes for the noun phrase as a whole. The grammatical attributes marked for the whole phrase include: part of speech, countability, collectivity, gender, verb agreement and article usage. The example given in the previous section, "traveling post office", was coded as a common countable noun that may be preceded by both the definite and indefinite articles and is referred to by the pronoun 'it'. Since the phrase "traveling post office" does not have any special requirements on verb agreement, that is, when the noun is used in the singular form it is followed by a singular verb and conversely, when it is used in the plural form it is followed by a plural form verb, the verb agreement marking is left blank for the entry.

4.3.2 Verifying Morphological Information

Although decisions for the descriptions of the intra-phrasal syntax structure were based on initial coding phases of the EDR Word Dictionary development, verification and correction of that morphological information during the verification project was essential. The coding of the syntactic information for the phrasal may affect the morphological information for the entry thus requiring the verification of morphological information as well. The decisions regarding the morphological information including segmentation and part of speech of the constituents could be made with more precision if the syntactic structure of the phrase was taken into consideration.

The basic principles of headword determination and segmentation are as follows:

(1) a headword unit should be determined on the basis of whether the phrasal expression comprises a single unit of meaning;

(2) phrasal headwords should be segmented into those constituents which are also found as single headwords in EDR's English Word Dictionary.

In view of the second basic principle, the part of speech of a phrasal constituent is decided according to the lexical part of speech consistent with the part of speech of the constituent as a single word entry in the dictionary. For example, nouns functioning as adjectives in phrases like "corn stalk" are coded as nouns and verbs modifying nouns as in the phrase "jam session" are coded as verbs.

The treatment of hyphenated words as single words or words which should be broken down into separate constituents is a significant segmentation issue. Hyphenated words which are used on their own in Standard English should be treated as single constituents and hyphenated words which are not used on their own should be broken down into separate constituents. In the examples below, the constituents are separated by the slash (/) notation.

"X-ray/ /spectroscopy"
"deep-sea/ /angler"

"directed/-/energy/ /weapon"
"Bose/-/Einstein/ /statistics"

A decision on segmentation for some hyphenated words in phrasal entries is difficult to judge purely by intuition from looking at the individual phrasal entries. These types of entries have to be looked at as a whole with attention being given to wider usage, and in particular to consistency with other headwords in the dictionary. For example, a decision to correct "yellow/-/green" to "yellow-green" cannot be made purely by intuition. The decision here is more an issue of the selection of headwords rather than one of hyphenation. The verification task of morphological information also included raising possible additional headwords to be added to the EDR English Word Dictionary through the analysis of the entries. After a decision has been made regarding entering the hyphenated word of the phrasal as a headword in the dictionary, the phrasal is fed back to the segmentation process.

4.4 Some Results of the Work

4.4.1 Syntactic Patterns

The result of the coding shows that 98% of the 34,650 phrasal entries could be covered in approximately 40 different patterns. The following seven patterns are the most frequent and cover over 80% of the total entries.

# Entries	Pattern	Example
1. 17422	EN1()/EN1(@)	hammock chair
2. 10847	EAJ()/EN1(@)	blue jay
3. 993	*EN2()/EN1(@)	Doppler effect
4. 707	*EN1(@)/EPP(EPR()/EN1())	piece of cake
5. 456	EAJ(EVE()/EEV())/EN1(@)	circulating library
6. 326	EN1(EVE()/EEV())/EN1(@)	changing room
7. 294	*EN1(EN1()/EEN:ENPOS())/EN1(@)	teacher's pet

*EN1 denotes a common noun, EN2 denotes a proper noun, EEN:ENPOS a noun possessive ending 's and ', EPR a preposition, and EPP a prepositional phrase. The location of the head of the phrasal is indicated by the @ notation.

4.4.2 Grammatical Attributes of the Constituents

As mentioned earlier one of the tasks of the coding was to indicate the grammatical attributes of the constituents. The data show that of the adjective + noun pattern (EAJ()/EN1()), the adjective constituent of the noun phrase did not inflect to form the superlative or comparative degree forms, but rather most often occurred in the positive degree form.

The grammatical attributes for nouns other than the head noun also showed some interesting results. Nouns other than those designated as the head noun do not inflect in most of the cases. One of the exceptional cases is "the time of w#one's life," where "w#one's" is a word class name for any noun in the possessive form. In this example, "life" inflects in accordance with the content of "w#one's" word class, though it is not the head noun of the phrase. Since phrases like this are very rare, it is also possible to treat "the time of w#one's life" and "the time of w#one's lives" as individual headwords and not as the inflected forms of the same headword. Another exceptional case in which more than one constituent could inflect would be phrases containing the conjunction 'and.' However, most of the phrasal entries in the form of 'A and B' are uncountable and the final noun inflects if the phrase is countable, such as "gin and tonics."

Therefore, we can assume that the grammatical behavior of constituents of noun phrase entries can be properly described by indicating the head noun and coding the inflection information and grammatical attributes of the head noun.

4.4.3 Grammatical Attributes for the Noun Phrase Unit

The coding of grammatical attributes for the entire noun phrase unit also provided some interesting results on countability and the usage of articles with the noun phrase. As is expected, the most typical combination of countability shows the following combinations:

If the noun is countable it may be preceded by the definite article or the indefinite article; If the noun is uncountable it may be preceded by the definite article or no article.

Approximately 10% of the nouns coded as countable showed a variation on the forementioned pattern. These countable nouns were coded as allowing the definite article, indefinite article as well as no article. Nouns with this type of coding included mass nouns, names of plants and animals, metals, food, titles etc. or other nouns which could refer to both the group or a member of the group. Examples of such nouns included "Leconte's sparrow", "Madagascar jasmine", "assembler language", and "atomic weight". Though this held for the majority of these types of nouns, it was not universally applicable; the use of no article with "Nubian goat", "Oregon grape" and "arctic loon" is questionable.

The significance of this data is that it implies perhaps a new code is necessary to cover cases of countable (ENC) nouns becoming uncountable (ENU) nouns and vice versa. Instead of coding a single entry as both, or providing two entries which correspond to the ENC and ENU usage we might better express the grammatical behaviors which are commonly shared by particular types of nouns by using a new code.

4.4.4 Verification of Morphological Information

In the morphological data some entries of the original data were segmented into constituents and some were not. This was particularly the case with '-ing' and '-ed' forms of words. The segmentation was not always consistent. But through syntactic analysis, verification of the segmentation and part of speech assignment could be carried out.

The EDR English Word Dictionary does not contain gerunds or participle forms of a verb as separate headword entries (except for irregular inflected forms). If a word in the '-ing' form is regarded as a gerund or a present participle, it is to be segmented into a verb and a verb ending. There are some cases where gerund forms or participle forms have been accepted as lexical items and not as inflected forms of a verb. In such cases, they are identified not as verbs, but as nouns or adjectives.

Noun phrases consisting of a word in the '-ing' form and another noun are treated by using one of the following four patterns, where EVE denotes an English verb and EEV a verb ending:

(1) EN1()/EN1(@)
"hunting knife"

(2) EAJ()/EN1(@)
"flying fox"
"man-eating shark"

(3) EAJ(EVE()/EEV())/EN1(@)
"intervening sequence"
"circulating medium"

(4) EN1(EVE()/EEV())/EN1(@)
"changing room"
"participating insurance"

If a phrasal in the form of '-ing + noun' could be reworded as 'a noun that is v-ing' or 'a noun that v-s' the entry was coded using either pattern 2 or pattern 3.

Through the verification of the morphological information we were able to gain more consistency in the segmentation of the constituents of phrasal headwords. Also we were able to indicate possible additional headword entries through the verification of the constituents that comprise the phrasal.

5. Conclusion

The syntactic structure of noun phrasal entries is described in a relatively small number of patterns. By coding a large number of noun phrasal entries it is possible to obtain an exhaustive list of syntactic patterns for noun phrases that would be listed as headword entries for English dictionaries. By describing the syntactic structure it is possible to obtain the syntactic information which is necessary to identify the internal structure of the phrasal as well as confirm and improve upon the segmentation of constituents and part of speech assignment to each constituent of the phrasal entry.

The vast majority of additional vocabulary, not only in the EDR English Word Dictionary, but in dictionaries in general will most likely be noun phrases. By utilizing the results from the current improvement project, the list of syntactic patterns for noun phrase entries can be used to check the appropriateness of the phrases as dictionary headwords as well as provide screening in order to prevent the recording of ill-formed structures, and finally to indicate syntactic ambiguity in the noun phrase itself.

References

EDR Electronic Dictionary Technical Guide (EDR, 1993)

Procter, P. (1992) The Cambridge Language Survey. Cambridge; Cambridge University Press.

Suematsu, H., Sugiura, M., Arioka, M. (1992) "A Distributive Representational Framework for English Collocations in an Electronic Dictionary," in: Lingvisticae Investigationes, XVI:2. John Benjamins, Amsterdam. Pages 373-394.

Wilks, Y., Slator, B., Guthrie, L. (in press) Electric words: dictionaries, computers and meanings. Cambridge, MA. MIT Press.

Yokoi, T. (1990) Towards information technology. Kyoritsu Shuppan.