Part-Of-Speech Labelling and the Retrieval of Phraseological Units

Alenka Vrbinc
Slovenia Faculty of Economics

The paper presents some insights into the problems of PoS labelling of lemmata, paying special attention to locating phraseological units where it is essential to identify the correct part of speech of the word under which the phraseological unit is included and dealt with in monolingual learners' dictionaries. When studying the inclusion of individual words, senses and phraseological units in five learners' dictionaries (OALD7, LDOCE5, COBUILD5, CALD3, MED2), it was found that numerous lemmata are equipped with more than one PoS label. Consequently, the user no longer needs to identify each and every part of speech of the word in question. As far as the inclusion of phraseological units in the entry for a specific part of speech is concerned, another method of inclusion is proposed, which can be regarded as a further simplification of the microstructure: i.e., that all phraseological units with one common element belonging to different parts of speech are simply grouped together in one special idioms section without distinction between individual parts of speech. This method is certainly worth applying in monolingual learners' dictionaries.

1. Introduction

For a long time, learners' dictionaries have included grammatical information in their word entries to help users not only in decoding but also in encoding; however, they take a non-specialized approach to grammar. When relatively sophisticated grammatical information is included, it is simplified to meet the needs of users of learners' dictionaries. Grammatical information is generally presented in traditional terminology because users are assumed to be familiar with such terminology. One of the most basic types of grammatical information is the part-of-speech label. Besides assigning the syntactic role to the lemma, part-of-speech labels also play an important role in including run-on entries (especially phraseological units) in the microstructure of the dictionary, which means that their importance when trying to locate a phraseological unit is not negligible.

Parts of speech are a grammatical category dealt with in all grammar books, from elementary ones to those intended for more professional purposes and consulted by experts, such as linguists. Every child gets acquainted with this category at a very early stage of learning the rules of his/her mother tongue. But how is this knowledge reflected in practice? Almost all dictionaries available on the market today, whether monolingual or bilingual, base the internal arrangement of homonymous articles on the principle of parts of speech, thus presupposing that an average dictionary user will have no problem identifying the part of speech of the word in question. Most certainly, this is not always an easy task. We have to bear in mind that not all parts of speech are equally easy or difficult for an average dictionary user to recognize (e.g., verb and noun, on the one hand, and conjunction and preposition, on the other); moreover, there is a very specific vocabulary item, namely the phraseological unit, where determining the part of speech of individual constituent elements may pose a problem even to a professional.

Many studies of dictionary use have been conducted investigating the use of grammatical information included in learners' dictionaries. Questionnaires used as a standard method for testing users' abilities to interpret and apply actual grammatical information provided by learners' dictionaries include questions about transitivity, verb complementation, countability,

gradability, attributive/predicative use, abbreviated phrases and other grammatical codes (cf. McCorduck 1993, Atkins, Varantola 1998, Tono 2001, Bogaards, van der Kloot 2001, Vrbinc, Vrbinc 2006, 2007). In other studies, the criteria users take into consideration when trying to locate multi-word units have been examined (Béjoint 1981, Tono 2001, Bogaards 1990, 1992, Atkins, Varantola 1998). To my knowledge, no studies of dictionary use and users' skills concentrate specifically on the problem of part-of-speech labelling in relation to the inclusion of phraseological units, despite all the differences that can be observed in different monolingual dictionaries.

When studying the inclusion of individual words, senses and phraseological units in five leading British monolingual learners' dictionaries (OALD7, LDOCE5, COBUILD5, CALD3, MED2), it was found that numerous lemmata are equipped with more than one PoS label. It has to be stressed that the policy of multiple PoS labels is relatively new, since it can only be found in recent revised editions of these dictionaries. This observation refers particularly to OALD and LDOCE, which is logical, given that these two dictionaries have the longest tradition (OALD was first published in 1948, whereas the first edition of LDOCE was published in 1978). Previous editions tended to stick to the traditional labelling of parts of speech: one entry – one PoS label. Phraseological units, however, are included and treated under a separate part of speech according to the part of speech of the elements that constitute a phraseological unit. Another method is to treat different parts of speech within the same entry. In this case, phraseological units are included at the end of the entry (e.g., COBUILD). Apart from that, there are dictionaries (e.g., *Oxford Dictionary of English*, 2nd edition) that do make a distinction between different parts of speech but group phraseological units together regardless of the part of speech of the word that is a component element of a particular phraseological unit. Arrangement regarding the part of speech is also not observed in phraseological dictionaries, where phraseological units are listed together according to the form of the first lexical word in the phraseological unit.

The article presents some insights into the problems of PoS labelling of lemmata, paying special attention to locating phraseological units where it is essential to identify the correct part of speech of the word under which the phraseological unit is included and dealt with in monolingual dictionaries. The emphasis is on five British monolingual learners' dictionaries, since apart from bilingual dictionaries, these dictionaries are most frequently consulted by learners of English and generally by non-native speakers of English. The most problematic parts of speech are adjectives, adverbs, conjunctions and prepositions, since the determination of these parts of speech poses the most difficulty not only for learners of a foreign language but also for native speakers of a particular language.

2. Part-of-speech labelling and the resulting inclusion of phraseological units in monolingual dictionaries

A close consideration of subsequent editions of one dictionary shows that much has been done in recent decades to simplify the internal structure and especially the different types of information included in a dictionary. The traditional grammatical division of entries has been replaced by semantic division in some dictionaries (e.g., all editions of COBUILD, first and second editions of CALD), but a certain degree of grammar has also been preserved in these dictionaries. A feature typical of COBUILD5 is that the entries are sometimes separated

according to their part of speech (e.g., hand), but more often one entry includes different parts of speech (e.g., play). In the latter case, the senses of different parts of speech that share a certain semantic component are treated as subsequent senses of the entry word. The system of semantic division of the entries is different in CALD2 from that in COBUILD5. In this dictionary, entries are included on the basis of the semantic meaning and part of speech. For the word 'hand', for instance, CALD2 includes 10 entries, the first one labelled noun, the second verb and the rest being nominal entries. The very last entry for 'hand', which is a noun, is followed by the phrasal verbs section. The word 'play' has nine entries, six of them being verbs and three of them being nouns (the second, the fourth and the sixth entries). Phraseological units are also included in different entries according to the semantic meaning. This way of including phraseological units is far from being user-friendly, since the semantic meaning of individual words in phraseological units is very often hard to define. In the last, i.e. the third, edition of CALD, the system of treatment of entries as well as phraseological units is completely different and closely resembles the traditional arrangement that can also be found in OALD7 and MED2. Interestingly, the electronic version of CALD3 has retained the semantic arrangement of entries, but the phraseological units are included at the end of the last entry of the appropriate part of speech.

An interesting system of including phraseological units can be seen in the *Oxford Dictionary of English*. The lexicographers did not make a distinction between parts of speech of individual words within a phraseological unit, but rather included all phraseological units after the last entry. For example, the word 'hand' has two entries (noun, verb) but the phraseological units follow the second of the two entries, i.e. the verb, disregarding the part of speech of the word 'hand' in individual phraseological units. This is by far the most user-friendly system of including phraseological units in dictionaries because no demands whatsoever are imposed on the user. It should be stressed that this dictionary is intended for native speakers of English, in contrast to the dictionaries that are taken into consideration in this contribution (learners' dictionaries). Consequently, we can say that the system of including phraseological units is simpler and more user-friendly in a dictionary that is not intended for learners of English.

3. Different PoS labelling in monolingual learners' dictionaries

On the basis of the analysis of PoS labelling of different lemmata, we could identify the following two groups in which problems concerning PoS labelling were detected:

(1) A single PoS label in each individual dictionary investigated, but different labels are found in different dictionaries. The combinations of parts of speech are numerous, thus this group can further be subdivided into the following subgroups:

a) adjective vs. adverb:
Of all the different parts of speech appearing in my research, this group was most numerous. For example:

do sb proud is listed under 'proud' in OALD7, LDOCE5, CALD3 and MED2 (it is not included in COBUILD5). However, the dictionaries do not agree on the part of speech of the entry word 'proud' and consequently not on the syntactic function of

'proud' in this particular phraseological unit. OALD7 lists this phraseological unit under the adverb, whereas LDOCE5, CALD3 and MED2 label 'proud' as an adjective. 'Proud' in 'do sb proud' is an adjective, since its meaning can be interpreted as 'do sth that makes sb proud of you'. Similarly, 'proud' in the old-fashioned meaning of this phraseological unit ('treat sb very well by giving them a lot of food, entertainment, etc.') can also be considered an adjective.

sooner or later is an interesting example that should be mentioned in connection with the confusion between adjectives and adverbs. In OALD7, it is treated under 'soon' as an adverb but listed with an accompanying cross-reference under 'late' as an adjective. In all other dictionaries, it is included and treated under 'soon' or 'sooner' as an adverb, but without a cross-reference at 'late' or 'later'.

b) adjective vs. noun:

keep mum and *mum's the word* are two phraseological units that share one component, namely the word 'mum'. When looking these two phraseological units up in learners' dictionaries, a dictionary user is faced with an intriguing situation: in OALD7, they are both included under 'mum' as an adjective; in MED2, they are included under 'mum' as a noun and in LDOCE5 and CALD3, 'mum's the word' is included under 'mum' as a noun and 'keep mum' under 'mum' as an adjective. To include 'mum's the word' under 'mum' as an adjective is an oversimplification, since in this phraseological unit 'mum' can be analysed as the subject complement and 'word' as the subject, the emphasis thus being on 'mum' (The word is mum. >> the emphasis on 'mum' >> inversion: MUM's the word.). In 'keep mum', the word 'mum' is syntactically an adjective, since 'mum' in this case refers to 'keeping information to oneself; silent', which is adjectival in meaning.

the long and (the) short of it is listed in all the dictionaries studied (except in COBUILD5, where it is not included) under the entry 'long'. The latter, however, is differently labelled as an adjective (in OALD7, LDOCE5) and as a noun (in MED2, CALD3). It seems reasonable to list it under the nominal entry, since 'long' is preceded by the definite article, which is a sure sign that the word following it is a noun.

c) adjective vs. verb:

rid is an interesting example, especially the phraseological units *be rid of sb/sth* and *get rid of sb/sth*. On closer examination, it can be seen that only OALD7 includes them under the entry labelled as a verb (the idioms section being followed by the phrasal verbs section containing the phrasal verbs 'rid sb/sth of sb/sth' and 'rid yourself of sb/sth'). In MED2, CALD3 and LDOCE5, both phraseological units are included under the adjectival headword, whereas in COBUILD5, the phraseological unit 'be rid of sb/sth' is labelled as an adjective, but the phraseological unit 'get rid of sb/sth' is labelled as being a phrase. It is correct to include these phraseological units under 'rid' as an adjective. 'Rid' would be interpreted as a verb if it were the

past participle of 'rid'. Despite its adjectival meaning, it would still be considered as a verbal form, which is not the case in these two phraseological units.

d) miscellaneous:

go phut is included in LDOCE5, CALD3 and MED2 (but not in OALD7 and COBUILD5). In all three dictionaries, we can find this phraseological unit under 'phut' as a noun, but syntactically speaking, 'phut' in this idiomatic expression is an adjective, since the verb 'go' in this phraseological unit is a linking verb, thus being followed by a subject complement. It can be claimed that 'phut' is a noun from the point of view of etymology but used in the adjectival sense in this phraseological unit.

The lemma *unquote*, which is a component element of the phraseological unit 'quote (... unquote)', is labelled as a noun in OALD7 (the cross-reference guides the user to the entry 'quote', which is labelled as a verb), MED2, CALD3 and LDOCE5 do not provide a PoS label for 'unquote', but they all include the phraseological unit 'quote ... unquote' in the entry for the verb 'quote'. COBUILD5 simply uses the explanation 'phrase'. It is interesting to mention two monolingual dictionaries for native speakers of English that are beyond the scope of this article: *Oxford English Dictionary* (2nd edn.) and *Collins English Dictionary* (9th edn.). The former lists the above phraseological unit under 'quote' as a verb, and a cross-reference is provided in the entry for the verb 'unquote'; in contrast the latter includes it under 'quote' as an interjection and label 'unquote' as used in this phraseological unit also as an interjection.

(2) A single PoS label in some dictionaries and a multiple PoS label in other dictionaries:

a) noun or adjective or adverb vs. conjunction, adverb:

When examining the lexical item *before long*, we can see that it is included in all five dictionaries examined, but the treatment varies widely from dictionary to dictionary. It should be mentioned that it is treated as an example of use in OALD7 but is included in two places in the dictionary and under two different entries: under the entry 'before' used as a conjunction (sense 1: 'We'll know before long (= soon).) and under the entry 'long' used as an adverb (sense 2: We'll be home before long (= soon).) In MED2, COBUILD5, CALD3 and LDOCE5, the same unit is treated as a phraseological unit but under different parts of speech. In MED2, it is included under the noun 'long', in CALD3 under the adjective 'long', in LDOCE5 under the adverb 'long' and in COBUILD5, the entry under which phraseological units are treated does not contain any PoS label but only the label 'phrases'. In LDOCE5, however, the same lexical item can also be found as an example of use in the entry for the preposition 'before' (sense 1: Other students joined in the protest, and before long (= soon) there was a crowd of 200 or so.). The above phraseological unit should be included in the entry for either the preposition 'before' or the noun 'long'.

b) adverb vs. adjective/adverb, determiner, pronoun:

The phraseological unit *at least* is treated under the entry 'least', which is variously labelled in different dictionaries. In OALD7, it is labelled as an adverb; in MED2, it is labelled as an adjective, adverb, pronoun and determiner (one entry including different parts of speech); in CALD3, 'least' is labelled as being an adverb, determiner and pronoun; in LDOCE5, it is a determiner and pronoun, whereas in COBUILD5, the part of speech is not given at all.

A similar example, but even more complicated, is 'little' as used in the phraseological units *little by little* and *as little as possible*. In OALD7, both units are included in the entry labelled determiner, pronoun, 'as little as possible' as an example of use and 'little by little' as an idiom. In MED2, both units have the same status as in OALD7, but the entry under which they are included is marked with a multiple PoS label (adverb, determiner, pronoun). LDOCE5 includes 'little by little' under the adverbial lemma as a phraseological unit, whereas 'as little as possible' is treated as an example of use under the lemma marked as a determiner or pronoun. In CALD3, the treatment is similar to that in LDOCE5, the only difference being that the entry under which 'as little as possible' is included is marked pronoun, noun. In COBUILD5, 'little by little' is marked 'phrase' and is included under the lemma labelled as 'determiner, quantifier, and adverb uses'.

c) adverb or conjunction vs. conjunction, determiner, pronoun, (adverb):

An example of multiple PoS labels of this kind is the unit *neither ... nor*. In OALD7, it is treated as sense 2 under 'neither' labelled as an adverb. In MED2, it is included as a phraseological unit in the idioms section under the headword labelled as a conjunction, determiner or pronoun. It is treated in a similar way in CALD3, the only difference being that apart from the previously mentioned PoS labels (conjunction, determiner or pronoun), it gives an additional PoS label: adverb. In LDOCE5, this phraseological unit is treated as such in the entry for 'neither' as a conjunction as well as under 'nor' as a conjunction/adverb and in COBUILD5 under 'neither' as a conjunction as an example of use.

4. Conclusion

It has to be pointed out that learners' dictionaries have simplified their microstructure to a great extent in the last couple of decades, PoS labelling and the inclusion of phraseological units being no exception. The problem of PoS labelling has been partly resolved by the introduction of multiple PoS labels; consequently, the user no longer needs to identify each and every part of speech of the word in question. As far as the inclusion of phraseological units in the entry for a specific part of speech is concerned, another method of inclusion should be mentioned, which can be regarded as a further simplification of the microstructure: i.e., that all phraseological units with one common element belonging to different parts of speech are simply grouped together in one special idioms section without distinction between individual parts of speech. Due to its user-friendliness, this method is certainly worth applying in monolingual learners' dictionaries.

More attention should be paid to the use of monolingual dictionaries at school because, after all, dictionaries are complex reference tools that are not easy to use if we want to extract all the information they contain. Despite all the simplifications observed in monolingual dictionaries in general, they still remain quite complicated for an average dictionary user.

References

Dictionaries

Cambridge Advanced Learner's Dictionary. 2nd edn. Cambridge: Cambridge University Press. 2005.
Cambridge Advanced Learner's Dictionary. 3rd edn. Cambridge: Cambridge University Press. 2008.
Collins COBUILD Advanced Learner's English Dictionary. 5th edn. London: HarperCollins Publishers. 2006.
Longman Dictionary of Contemporary English. 5th edn. Harlow, Essex: Pearson Education Limited. 2009.
Macmillan English Dictionary for Advanced Learners. 2nd edn. Oxford: Macmillan Education. 2007.
Oxford Dictionary of English. 2nd edn. Oxford: Oxford University Press. 2003.
Oxford Advanced Learner's Dictionary of Current English. 7th edn. Oxford: Oxford University Press. 2005.

Other literature

Atkins, S. B. T.; Varantola, K. (1998). 'Language learners using dictionaries: The final report on the EURALEX/AILA research project on dictionary use'. In Atkins, S. B. T. (ed.). *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Lexicographica Series Maior 88, Tübingen: Max Niemeyer Verlag. 21–81.
Béjoint, H. (1981). 'The foreign student's use of monolingual English dictionaries: A study of language needs and reference skills'. In *Applied Linguistics* 2 (3). 207–222.
Bogaards, P. (1990). 'Où cherche-t-on dans le dictionnaire?' In *International Journal of Lexicography* 3 (2). 79–102.
Bogaards, P.; van der Kloot, W. A. (2001). 'The use of grammatical information in learners' dictionaries'. In *International Journal of Lexicography* 14 (2). 97–121.
McCorduck, E. S. (1993). *Grammatical Information in ESL Dictionaries*. Lexicographica Series Maior 48, Tübingen: Max Niemeyer Verlag.
Tono, Y. (2001). *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension*. Lexicographica Series Maior 106, Tübingen: Max Niemeyer Verlag.
Vrbinc, A.; Vrbinc, M. (2007). 'EFL students' use of grammatical information in five leading British learners' dictionaries'. In *Linguistica* 47. 145–158.
Vrbinc, M.; Vrbinc, A. (2006). 'A research-based study of foreign students' use of grammatical codes in five leading British learners' dictionaries'. In *Linguistica* 46 (2). 227–242.