Inclusion Strategies for Multi-word Units in Monolingual Dictionaries*

Phillip Louw, *Department of Afrikaans and Dutch, University of Stellenbosch, Stellenbosch, Republic of South Africa (pal@sun.ac.za)*

Abstract: This article focuses on inclusion strategies for different types of multi-word units, be it as part of the macrostructure or embedded as treatment units in the microstructure of a specific dictionary. The types of multi-word units discussed range from multi-word lexical items to collocations and multi-word compound lexical items. The general principles set out in this article are applied specifically to monolingual school dictionaries that target learners of English in the junior secondary phase.

In order to discuss inclusion strategies adequately it is, however, necessary to make a cursory distinction between idioms and collocations, on the one hand, and between collocations and multi-word compound lexical items, on the other. It is shown that current monolingual dictionaries often fail to distinguish between these types and therefore apply potentially confusing inclusion strategies.

In the discussion of inclusion strategies for multi-word lexical items that follows, it is shown that, whereas loan groups and group prepositions require lemmatisation as full multilexical lemmas, the strategy for idioms is not as simple. The problems with a full lemmatisation of idioms are pointed out and an alternative system, whereby idioms are consistently included as sublemmas with full microstructural treatment, is proposed.

Next it is shown that collocations do not have lexical item status and can therefore not be treated in the same way as multi-word lexical items. However, provision must be made that some collocations may need additional microstructural treatment addressed to them.

Lastly, inclusion strategies for multi-word compound lexical items, which frequently occur in English, are discussed. The practice of sublemmatising so-called "transparent" compound lexical items and giving them no or little microstructural treatment, is shown to be inappropriate for school dictionaries.

Hopefully the guidelines provided in this article can be of some help in clearing up the muddled approaches currently followed in some South African monolingual school dictionaries.

Keywords: COLLOCATIONS, COMPOUND NOUNS, GROUP PREPOSITIONS, IDIOMS, INCLUSION STRATEGIES, LEMMA, LOAN GROUPS, MACROSTRUCTURE, MICROSTRUCTURE, MULTI-WORD COMPOUND LEXICAL ITEMS, MULTI-WORD UNITS, SUBLEMMA, TRANSPARENCY

Opsomming: Opnamestrategieë vir veelwoordige eenhede in eentalige woordeboeke. Hierdie artikel fokus op opnamestrategieë vir verskillende veelwoordige een-

* This article is an adaptation of a part of a chapter from a D.Litt. Dissertation *Criteria for a Multifunctional, Monolingual Dictionary in Junior Secondary Education*, which was accepted by the University of Stellenbosch , Stellenbosch, Republic of South Africa in April 2004.

hede, hetsy as deel van die makrostruktuur of ingebed as behandelingseenhede in die mikrostruktuur van 'n bepaalde woordeboek. Die tipes veelwoordige eenhede wat bespreek word, wissel van veelwoordige leksikale items tot kollokasies en veelwoordige samestellings. Die algemene beginsels wat in hierdie artikel uiteengesit word, word spesifiek op eentalige skoolwoordeboeke wat leerders van Engels in die junior sekondêre fase teiken, toegepas.

Om opnamestrategieë toereikend te bespreek, is dit egter nodig om eers 'n onderskeid te tref tussen idiome en kollokasies aan die een kant, en tussen kollokasies en veelwoordige samestellings aan die ander. Daar word aangetoon dat die huidige eentalige woordeboeke dikwels nalaat om tussen hierdie tipes te onderskei en derhalwe potensieel verwarrende opnamestrategieë toepas.

In die bespreking van opnamestrategieë wat daarop volg, word aangetoon dat, terwyl leenwoordgroepe en groepvoorsetsels lemmatisering as volle multileksikale lemmas benodig, die strategie vir idiome nie so eenvoudig is nie. Die probleme met 'n volle lemmatisering van idiome word uitgewys en 'n alternatiewe stelsel waardeur idiome konsekwent as sublemmas met volle mikrostrukturele behandeling opgeneem word, word voorgestel.

Vervolgens word aangetoon dat kollokasies nie leksikale-itemstatus het nie en dus nie op 'n soortgelyke wyse as meerwoordige leksikale items behandel kan word nie. Daar moet egter daarvoor voorsiening gemaak word dat sommige kollokasies wel addisionele mikrostrukturele behandeling mag benodig.

Laastens word opnamestrategieë vir meerwoordige samestellings, wat dikwels in Engels voorkom, bespreek. Die gewoonte om sogenaamde "deursigtige" samestellings te sublemmatiseer en dan van min of geen mikrostrukturele behandeling te voorsien nie, word as onvanpas vir skoolwoordeboeke getoon.

Hopelik kan die riglyne wat in hierdie artikel verskaf word, van hulp wees om die verwarde benaderings wat tans in sommige Suid-Afrikaanse eentalige woordeboeke gevolg word, op te klaar.

Sleutelwoorde: DEURSIGTIGHEID, GROEPVOORSETSELS, IDIOME, KOLLOKASIES, LEENWOORDGROEPE, LEMMA, MAKROSTRUKTUUR, MEERWOORDIGE EENHEDE, MEERWOORDIGE SAMESTELLINGS, MIKROSTRUKTUUR, NAAMWOORDSAMESTELLINGS, OPNAMESTRATEGIEË, SUBLEMMA

Introduction

Multi-word units present many problems to practical lexicographers, ranging from criteria for their selection through to the actual microstructural treatment afforded to each type. In this article the focus is, however, on inclusion strategies for different types of multi-word units, be it as part of the macrostructure or embedded as treatment units in the microstructure of a specific dictionary. The general principles set out in this article are applied specifically to monolingual school dictionaries that target both mother tongue and non-mother tongue learners of English in the junior secondary or senior phase. The types of multi-word units discussed range from multi-word lexical items (with specific reference to idioms), to collocations and multi-word compound lexical items (with specific reference to compound nouns).

Multi-word lexical items

Multi-word lexical items should be considered for inclusion in a monolingual school dictionary and can be lemmatised as multilexical lemmas. Gouws (1989: 97) states that "as a lexical item a multilexical lemma represents a single semantic unit, and the meaning of this unit cannot be deduced from the sum of the meaning of the constituent parts" [my translation]. Zgusta (1971: 154) adds that "for the lexicographer, the detection and correct presentation of multi-word lexical units is one of his most important tasks". Yet, what this correct presentation should be is a polemic issue. It may depend on the type of dictionary and may even differ for different types of multilexical lemmas, as illustrated in the following discussion. On a macrostructural level the lexicographer needs to decide whether these items should be lemmatised as main lemmas, or whether they can be listed under the first prominent constituent of the multi-word lexical item. Should the latter option be preferred, methods should be found not to perpetuate the confusing practice of grouping multi-word lexical items with collocations and examples.

Loan groups

Loan groups are perhaps the multi-word lexical items most consistently lemmatised as main lemmas. They are lemmatised in full and the space between constituent parts simply ignored when determining their place in the dictionary's sort order. This is also the practice in the *Chambers-Macmillan South African Dictionary Junior Secondary* (henceforth SADJS) and *The South African Oxford School Dictionary* (henceforth SAOSD), where loan groups such as **et cetera** and **post mortem** are lemmatised in full.

Group prepositions

The lexicographic treatment of group prepositions is a more disputed matter. Generally, the status of group prepositions as lexical items is not fully recognised in current monolingual standard and school dictionaries. They are often included as collocations and not as lemmas. Furthermore, even in dictionaries where their value and search priority are recognised and they are given sublemmatic status, group prepositions are often not sublemmatised under their first constituent parts.

If the group preposition **in aid of** is taken as an example, it soon becomes obvious that lexicographers identify **aid** as its main element and accordingly use that lemma as point of inclusion. This is the case in both SAOSD and SADJS where **in aid of** is given as a sublemma under **aid**. This practice unfortunately leads to the disruption of the initial alphabetical ordering principle adhered to elsewhere in these dictionaries. Such a disruption may be justifiable

in the case of idioms, where the dictionary culture leads users to look up idioms under the first prominent constituent. In the case of group prepositions, though, it is uncertain whether the dictionary culture dictates this practice to the same measure. It could therefore be contended that the users of a school dictionary may well be better served by listing group prepositions consistently as multilexical main lemmas. This would also be a lexicologically sound lexicographic practice.

Idioms

Should the model employed for other multi-word lexical items be perpetuated, idioms should also be lemmatised as multilexical main lemmas. However, this is not a practical solution for a school dictionary. Firstly, it will not always be possible to identify the initial component of the idiom. Articles are often interchangeable or optional at the start of an idiom and other subtle variations can occur. Secondly, lemmatisation of idioms can take up more space than alternative methods. Thirdly, the current dictionary culture (perpetuated by the available dictionaries) is one in which users of school dictionaries will probably expect to find the idiom as a sublemma under the lemma corresponding to the first word in the idiom that is considered to be semantically prominent, especially a noun, verb or adjective. This practice probably stems from the assumption that words function as independent lexical items in an idiom, rather than as constituent parts of an encompassing multi-word lexical item.

A case could be made out that a school dictionary is the ideal place to start changing the dictionary culture in subtle ways and that the lemmatisation of idioms should therefore be considered, as it is a lexicologically and lexicographically sounder method. However, the practice of including idioms as sublemmas is so strongly entrenched that such a move may be experienced as too unconventional and therefore result in users not finding the data they are looking for. Furthermore, the standard dictionaries, which these users are likely to use when school dictionaries no longer meet their needs, also predominantly give idioms as sublemmas, one of the functions of the junior secondary school dictionary being to prepare its users for a seamless transition to standard dictionaries. It may, therefore, be more advisable to work within these confines by ensuring that idioms are clearly marked and that the microstructural treatment of these idioms is as user-friendly as possible.

SAOSD and SADJS have opted to conform to the often-used practice of including idioms as sublemmas. There is, however, a problem with their approach. Should idioms be included as sublemmas, they need to be clearly distinguishable from microstructural data categories. This is unfortunately not the case in SADJS and SAOSD. Multi-word compound lexical items, idioms and collocations are treated similarly, making it difficult for the user to discern between these data types. The lemmatisation of multi-word compound lexical items will help to alleviate this situation, but the problem of possible confusion

between collocations and idioms remains. SAOSD bears witness to this confusion at the lemma **stick**, where the idiom **stick up for** is presented in exactly the same way as the collocations **stick out** and **stick to**. All three are included as sublemmas and given near-full microstructural treatment. As neither the collocations nor the idiom are deemed as fully transparent, items giving the meaning description are provided throughout.

Gouws (1996: 5) proposes the following solution:

> Belonging to separate information categories the collocations and idioms … should be accommodated in different article positions which will leave the user with different search areas allocated to each information category. By using different typefaces or structural markers the user could be lead to a clear distinction between these two information categories.

This solution can be modified in that idioms should rather maintain their lexical item status and function as sublemmas instead of entries within data categories in the microstructure. Collocations, on the other hand, will fit into that part of the comment or subcomment on semantics reserved for examples, but the possibility should still be there for less transparent collocations to be treatment units. The compiler(s) of a dictionary can also consider using an explicit structural indicator to show the start of the idiom group, as is practiced in WAT and HAT, for example, to ensure swift access. In terms of micro-architecture, it would also be advisable that each idiom, as well as the structural indicator introducing the idioms, start on a new line.

Collocations

Collocations are lexical combinations usually included in the microstructure as co-text entries in order to illustrate, what Gouws (1989: 227) refers to as "the typical microsyntactic context of the lemma" [my translation]. These combinations are typical and usually transparent. They therefore do not have lexical status as a whole, but comprehensive inclusion is still a necessity, "especially in pedagogical and translation dictionaries" [my translation] (Gouws 1989: 227). Cop (1991: 2776) states as reason for their inclusion that "even transparent collocations must be present, because they are not predictable". This sentiment is echoed by Svénsen (1993: 101): "Information about collocations is important in both monolingual and active bilingual dictionaries, since the user cannot be expected to know which words customarily occur together." Data on collocations provides microsyntactic empowerment, especially to users employing their dictionaries in an encoding task.

Transparency is, however, a problematic concept as users' perceptions of what is and is not transparent can differ greatly. It would therefore be wise for the lexicographer to err on the side of caution and ensure that collocations of which the transparency is at all doubtful, be included as treatment units. The

extent of the treatment will depend on the perceived lack of transparency (the lexicographer has to exercise sound judgement, but empirical research could also be of value here). It can include a short item giving a paraphrase of meaning, constructed examples showing the macrosyntactic use of the collocation, or a combination of these two data types.

If collocations are to be truly user-friendly, they "must reflect natural language" [my translation] (Gouws 1989: 227). It is therefore important that corpus data is analysed in order to identify possible collocations. The superior sorting abilities of the new generation of corpus-querying tools make this a more or less standard task for the lexicographer. These programs have the additional advantage of indicating to the lexicographer the frequency of use of each collocation. Should there then be a need to only select certain collocations, due to there being too many to include, the lexicographer can choose the most typical ones.

There is a marked difference in the treatment of collocations in SADJS and SAOSD. SADJS presents detailed example material in the form of collocations and constructed examples, whereas SAOSD opts for a larger macrostructure at the cost of linguistic examples. SAOSD does present some collocations and a very limited number of constructed example sentences.

There is, however, a significant problem in both these dictionaries' placement of less transparent collocations acting as treatment units. As has been mentioned in the discussion of idioms, these collocations are displaced from the normal search zone for syntactic data and moved to the end of the article to be lumped together with compound nouns consisting of more than one constituent, and idioms. This move has various implications. Firstly, it is very difficult for the target user to determine which type of data is being dealt with. This treatment therefore clashes with a basic lexicographic principle, i.e. that each data type should be treated distinctly. Secondly, the displacement disrupts the coherence in the search zone for examples, in that the micro- and macrosyntactic data are now distant.

Lastly, and perhaps most importantly, SAOSD and SADJS's placement of collocations at the end of an article complicates the search path at polysemous lemmas or lemmas with more than one syntactic function. Collocations may vary in transparency, but the guiding principle in determining whether a phrase is a collocation is still that there should be a discernable correspondence between the lexical item represented by the lemma's manifestation in the collocation and the meaning of either the lexical item represented by the lemma or the sense of the lexical item represented by the lemma. At **long** in SADJS, for example, the collocation **before long** corresponds to the second sense of **long**, but interspersed between them are the third sense, a compound noun (**long jump**) and an idiom. The user has to follow a complicated remote addressing procedure to bring all the relevant data together. It would therefore be much more sensible to include the collocation at the specific sense or syntactic function it corresponds to.

Compound lexical items

Besides the fundamental distinction between idioms and collocations, there is also another necessary distinction, i.e. between collocations (as microstructural items) and multi-word compound lexical items (as macrostructural items). The inclusion of compound lexical items provides another difficult macrostructural challenge to the compiler of a monolingual English dictionary. This challenge is specifically rooted in the variation in spelling of these compound lexical items, the appropriateness of sublemmatisation to the target user group and the question of transparency. In the following paragraphs, the treatment of compound nouns is used as an example, as this category best illustrates the variety of problems faced in the macrostructural treatment of compound lexical items.

Béjoint (1999: 81) comments that using "graphic cohesion" as a criterion to distinguish compound nouns "is difficult to apply, particularly in English, because of the variations in spelling: an English compound noun like *paper clip* can have the forms XY, X Y, or X-Y". He (1999: 82) adds that "this makes the automatic extraction of compounds particularly difficult in English". The lexicographer should obviously not have trouble identifying one-word or hyphenated compounds and considering them for lemmatisation or sublemmatisation, but the so-called "open compounds" (written as two words) can be a more challenging prospect. As is hinted at by Béjoint, it is often difficult to extract these from corpora, especially without having sophisticated software with corpus-querying tools that can sort according to context on the right of the search term. This problem underlines, once again, the need for such software. Furthermore, it can be very difficult to determine whether the combination dealt with is a compound noun or a collocation. Here the lexicographer's intuition, as an advanced language user, will play an important role, but more scientific criteria can be identified to aid in the task. Béjoint (1999: 82) lists some of these criteria as "non-compositionality", "position of the stress", "frequency", and "lexical unity". (For a more detailed discussion of these criteria, see Béjoint 1999: 82.)

Once a method of distinguishing between collocations and multi-word compound nouns has been found, the treatment of this type of compound noun can be contemplated. These compound nouns should be treated in the same way as single-word or hyphenated compound nouns, because "despite the blank, these compounds will be identified as one concept and therefore one base form …" (Schnorr 1991: 2815). The question then arises whether compound nouns should be included as lemmas or as sublemmas.

The use of sublemmas can be an important space-saving mechanism when they are given a limited microstructural treatment (e.g. just part-of-speech indication). Yet, there are serious reservations regarding their appropriateness for use in a school dictionary, which must be addressed by any prospective compiler. Firstly, either nesting or niching must be identified as the user-friendliest ordering method. Secondly, the compiler(s) must discern which types of mor-

phologically complex items can be sublemmatised. Thirdly, the level of textual condensation of the specific sublemma signs must be addressed, as, for example, omitting the part of the sublemma that corresponds to the lemma can save space, but could also alienate the target user group if it does not understand this procedure. Compound nouns present a particularly taxing task to the lexicographer when considered for sublemmatisation.

In many dictionary projects, the decision to lemmatise or sublemmatise depends on the level of transparency of the compound noun. Should a compound noun be deemed transparent, i.e. that the sum of the meanings of its constituent parts is equal to the meaning of the whole, it is often sublemmatised and given a limited microstructural treatment. Transparency is, however, a highly subjective criterion that requires the lexicographer to make assumptions as to which compound nouns the target users of the dictionary may experience as transparent. Béjoint (1999: 84) correctly surmises that "the actual transparency of a compound noun varies according to the ability of each language user to understand its elements."

The assumption of transparency is particularly difficult in the compilation of a junior secondary school dictionary, as there is a considerable difference between the linguistic skills and intuition of the lexicographer and the target user group. Furthermore, the target user group is a diverse group with great variation anticipated in the linguistic skills and intuition of its individual members. To this dilemma could be added that, even if the constituent parts of a compound noun are recognised by the user, confusion could still arise as to which senses of the constituents are activated by their functions in the whole. A good case can therefore be made out that the compiler(s) of a dictionary should not readily assume transparency, but rather give a full microstructural treatment to each compound noun that meets the frequency requirements for inclusion. The lemmatisation of compound nouns, as guiding elements of default single articles, would be one way of achieving this goal.

Both SADJS and SAOSD have a somewhat unusual approach to the treatment of compound nouns. In both dictionaries single-word or hyphenated compound nouns are usually lemmatised, whereas those of the multi-word variety are sublemmatised, but, strangely, given detailed microstructural treatment. Examples of this type of treatment are provided by **grandfather clock** (sublemmatised under **grandfather** in both SADJS and SAOSD), **further education** (sublemmatised under **further** in SADJS) and **rat race** (sublemmatised under **rat** in SAOSD). This treatment is not always applied consistently. SAOSD, for example, gives **sulphuric acid** full lemma status, whereas it could (following similar examples in SAOSD) have been sublemmatised under **sulphur**.

The sublemmatisation procedures for compound nouns followed in SADJS and SAOSD can be motivated from the point of view that it places these items where they morphologically belong. More research needs, however, to be done to determine whether the target users of the dictionary would expect to find

these items as sublemmas. It may be contended that giving these items full lemma status and ignoring the space between constituent parts when ordering them in with single-word lemmas would better meet the user expectations and be less confusing. It would also solve the problems presented by decisions based on perceived transparency.

Conclusion

Hopefully the guidelines provided in this article can be of some help in clearing up the muddled approaches currently followed in some South African monolingual dictionaries. Such improvements will, in turn, be to the benefit of the users of these dictionaries, especially those whose linguistic ability is still at a formative stage.

Bibliography

Dictionaries

Chambers-Macmillan. 1996. *Chambers-Macmillan South African Dictionary Junior Secondary*. Manzini: Macmillan Boleswa Publishers.

Hawkins, J.M. 1996. *The South African Oxford School Dictionary*. Cape Town: Oxford University Press.

Other sources

Béjoint, H. 1999. Compound Nouns in Learners' Dictionaries. Herbst, T. and K. Popp (Eds.). 1999: 81-99.

Cop, M. 1991. Collocations in the Bilingual Dictionary. Hausmann, F.J. et al. (Eds.). 1989–1991: 2775-2778.

Gouws, R.H. 1989. *Leksikografie*. Cape Town: Academica.

Gouws, R.H. 1996. Idioms and Collocations in Bilingual Dictionaries and their Afrikaans Translation Equivalents. *Lexicographica* 12: 54-88.

Hausmann, F.J. et al. (Eds.). 1989–1991. *Wörterbücher. Ein internationales Handbuch zur Lexikographie/Dictionaries. An International Encyclopedia of Lexicography/Dictionnaires. Encyclopédie internationale de lexicographie*. Berlin/New York: Walter de Gruyter.

Herbst, T. and K. Popp (Eds.). 1999. *The Perfect Learners' Dictionary (?).* Lexicographica. Series Maior 95. Tübingen: Max Niemeyer Verlag.

Schnorr, V. 1991. Problems of Lemmatization in the Bilingual Dictionary. Hausmann, F.J. et al. (Eds.). 1989–1991: 2813-2817.

Svensén, Bo. 1993. *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford: Oxford University Press.

Zgusta, L. 1971. *Manual of Lexicography*. Prague: Academia / The Hague/Paris: Mouton.