The codification of phonological, morphological, and syntactic information

Geert Booij (Vrije Universiteit Amsterdam)

1. Introduction

It is quite obvious that an adequate monolingual dictionary must be based on large electronic corpora in order to function as a reliable guide. The dictionary itself should also have an electronic form, from which a printed form can be derived. The electronic form is not only essential in the production of a dictionary (updating, consistency checks, etc.), but also for the user: in combination with adequate search programs, an electronic dictionary provides far more information than can be found by means of consulting an alphabetically ordered list of lexical items, the traditional form of dictionaries in printed form. For instance, an electronic dictionary makes it very simple to find all words that contain a particular letter sequence, and thus can function as a research tool for phonologists and morphologists.

If a dictionary is based on large and representative corpora, it is also possible to provide reliable data on frequency use. This kind of information is important for developing adequate study materials for first and second language acquisition, including training in orthography, and may also be employed by the dictionary user to infer the status of a word: is it a common word, or rather obsolete?

A final preliminary remark is that a good dictionary should also be based on both written language corpora and spoken language corpora because there are many words that are characteristic of spoken language, or, conversely, occur in written language only. Traditional dictionaries tend to be biased towards written language, but this can be corrected now that spoken language corpora are more and more becoming available. Without a good corpus of spoken language the lexicographer will easily forget to include words that are typical for spoken language.

2. Phonological information

The primary phonological information on each lexical item to be provided by the dictionary is its phonetic form. With 'phonetic form' I mean the phonetic form of the word as spoken in isolation, in careful speech. This phonetic form must be given in the notation of the International Phonetic Alphabet. There is often variation in the phonetic realisation of words, however. For instance, in Dutch the word *banaan* 'banana' is pronounced as [ba:'na:n] in isolation. It also has the phonetic forms [bΨna:n] and [bχna:n] in connected speech, due to the phonological processes of vowel shortening and vowel reduction respectively (Booij 1995). We might think that it makes no sense to include all these phonetic forms in the dictionary since they are predictable. However, it appears that this variation is (partially) lexically governed. For instance, of the Dutch words *minuut* 'minute' and *piloot* 'pilot', both with an /i/ in the first unstressed syllable, it is only in the first one that the /i/ can also be realised as schwa: high vowels are only reduced in words of relatively high frequency such as *minuut*. Hence, this information on the details of the phonetic realisation of words is lexical information, and should therefore be included in the dictionary. Vowel reduction is a good example of a phonological process that is subject to lexical diffusion: words are affected one by one by this process. Therefore, the outputs of such processes have to be listed in the dictionary.

The phonetic form should also be encoded acoustically, so that the dictionary user can hear the word being spoken by clicking on the phonetic form in the entry for that word. This feature of a good modern electronic dictionary has been made possible by present-day information technology, and is particularly useful for second language training.

The segmental composition of the phonetic form is not the only useful information. In addition, we should represent the information on the location of primary and secondary stress on the syllables of the word, and the division of the word into its syllables. Representation of stress location is certainly necessary for those languages for which it it is not fully predictable, such as English and Dutch. In languages such as Finnish, French, Polish, with regular stress, it will suffice to only represent stress on exceptional words (borrowings).

Information on syllabification is also useful, because it is not always fully predictable. This is illustrated by the Dutch word *aardappel* 'potato'. Originally, this word was a compound, with the constituents *aard* 'earth' and *appel* 'apple'. However, synchronically, it is no longer experienced as such, and hence it is syllabified a simplex word. In compounds, the internal morphological boundary coincides with a syllable boundary. Thus, the syllabification of *aardappel* changed from *aard.ap.pel* into *aar.dap.pel*, unlike that of the structurally identical compound *handappel* 'lit. hand apple, eating apple'. This example shows that syllabification may be lexically governed, and thus belong to the realm of lexical information.

In some languages (for instance, Dutch), the syllabification of a word strongly correlates with the possible hyphenation patterns of the orthographic forms, because the hyphens coincide with syllable boundaries. This is another reason why information on syllabification patterns is useful. Note, however, that syllabification and hyphenation do not always fully coincide. For instance, the word *aardappel* discussed above is hyphenated as *aard-appel*, that is, as if it is still a compound. Therefore, unpredictable and exceptional cases of syllable-based hyphenation must be represented in the lexicon.

English is different from Dutch in that English hyphenation reflects morphological structure, if possible, rather than phonological structure (compare the hyphenation of the Dutch adjective *a-gres-sief* to its English equivalent *aggress-ive*). This kind of hyphenation is even less predictable because it is often impossible to assign a straighforward morphological structure to an English word. Hence, lexical information about English hyphenation is even more necessary than in the case of Dutch, and is indeed given in most dictionaries of English.

The phonetic form of a morpheme may vary depending on the morphological context in which it occurs. For example, the Dutch lexical morpheme *hoed* 'hat' is pronounced as [hut] when it is used as a word in isolation, as a singular form, but as [hud] in the plural form *hoeden* [hudχn]. This allomorphy (= alternation in the phonetic shape of a morpheme) is not predictable, as is shown by the similar word *voet* 'foot' [vut] with the plural form *voeten* 'feet' with the phonetic [vutχn]. Therefore, the fact that *hoed* exhibits this alternation, is lexical information.

A standard way of representing this kind of information on alternation in present-day phonology is by making use of the notion 'underlying form'. For instance, we may assign the underlying phonological form /hud/ to *hoed.* When used as a singular form, without an additional vowel-initial suffix, the underlying /d/ appears in syllable-final position, and hence it is predictably realized as voiceless [t], due to the phonological constraint of Dutch that obstruents (stops and fricatives) are always voiceless at the end of a syllable. In the plural form, the morpheme-final /d/ begins the second syllable, and hence it is not subject to devoicing. In contrast, the underlying form of *voet* is /vut/, which implies that there is no alternation in the phonetic form of this morpheme.

As we saw, this lexical information can be expressed by giving the underlying phonological form of each word in its lexical entry. Alternatively, we might want to avoid making use of the theoretical notion 'underlying form', and represent (a subset of) the inflectional forms of a word, with their phonetic forms. It is then left to the dictionary user how to interpret such phonological variation in the set of inflectional forms. This second option is the better one in those cases where phonologists might disagree as how to account for allomorphy because there are always two options: assigning the allomorphs (the phonetic variants of a morpheme) a common underling form and deriving the phonetic forms by means of a set of rules or constraints, or listing the allomorphs of each word, with possibly additional statements about the distribution of the allomorphs. For instance, we might assign one common underlying form /sign/ to the part *sign-* of both *sign* and *signal*, and derive the different allomorphs of *sign* ([sajn] or [sign]) by rule from this underlying form, as in Chomsky & Halle (1968). Alternatively, we may assume two listed allomorphs for the lexical morpheme *sign*, and this is the preferred option in present-day phonological theory since it avoids a too abstract derivational analysis. The reason why this second option is better is because a dictionary should not be loaded with theory-dependent information that might easily change.

This position implies, as stated above, that we list inflectional forms of words in the dictionary, but as we will see below, there are independent reasons for doing this.

3. Morphological information

A standard assumption about morphological information in dictionaries is that regular inflectional morphology need not be specified in the dictionary. For instance, we may take the position that the different inflectional forms of verbs in Germanic languages need not be specified if the verb is regular, but only when it has irregular forms, as is the case for the past tense and participle forms of the so-called strong or stem-alternating verbs. In this respect, inflection receives another treatment in the dictionary than word formation, for which the dictioanty also lists the regular forms. The reason for this difference is the treatment of regular morphology is that inflection deals with different forms of the same word (in the sense of 'lexeme'), whereas word formation is a matter of creating new lexemes.

Word formation processes (derivation, compounding, etc.) define the set of possible words of a language, but this is not enough: we need to know if a possible morphologically complex word actually exists. The dictionary thus provides information about the lexical conventions of a language. Hence, existing (‘established’) complex lexemes should be listed in the dictionary, either as separate lexical entries, or –in order to reduce the size of a dictionary - as part of the entry of their base word. In the latter case, compounds should be mentioned in the entry for their head, which is the right constituent in Germanic compounds. Thus, Dutch *handappel* 'eating apple' should be listed under *appel*, not under *hand*, because language users know that the word *handappel* stands for a subset of apples, not of hands, and will look for this word under the heading for *appel.*

Nevertheless, there are morphological arguments for including information on inflected forms of words in the lexicon: formal irregularity - as we saw above - and unpredictability. An illustration of unpredictability is that the pluralization of Dutch nouns is not fully predictable. Dutch has two plural suffixes, -*s* and -*en*. There is a division of labour between these two suffixes (basically, -*s* occurs after stems ending in an unstressed syllable, -*en* after stems ending in a stressed syllable, cf. Booij 2002a), but there is a large number of exceptions. For instance, loans from English take -*s* instead of the predicted -*en*, as in *flats* 'id.'. Some nouns have two plural forms, as is the case for *zoon* 'son' with the plural forms *zoons* and *zonen*. Hence, the plural forms of Dutch nouns should be listed in the dictionary. Moreover, in the case of pluralization of nouns, it appears that many of them do not have a plural form at all. This is also lexical information, and therefore, we have to list each individual existing plural form. As to the inflectional forms of verbs, the situation is slightly different. Normally, each verb has all forms for morphosyntactic categories such as person and number, so the issue whether a particular form exists, does not arise. That is, gaps in the verbal paradigm are more exceptional. Defective verbs do occur, however, and for such verbs it has to be specified which forms are available. For instance, Dutch has verbal compounds such as *hardlopen* 'to run fast' that do not have finite forms. As to the inflection of adjectives, it is necessary to list comparative and superlative forms since not all adjectives do not have them. For instance, intensifying adjectives such as *steenkoud* ‘ lit. stone-cild, very cold’ do not have degree forms: **steenkouder, steenkudst*. On the other hand, the inflection of adjectives as determined by agreement need normally not be listed, because we expect each adjective to have such a form. Yet, they have to be listed as part of the noun phrases in which they occir, if the inflectional form of the agreeing adjective is not fully predictable This is the case for *een goed mens* ‘a good human being’, in which *goed* ‘good’ lacks the suffix schwa that normally occurs in attributive position.

All kinds of inflection may exhibit allomorphy of the type discussed above (alternation between voiced and voicelses obstruents), which we might represent directly, by means of listing the phonetic forms of the inflected forms. That is, there might also be phonological reasons for listing the inflectional forms of a word in the dictionary. In sum, a dictionary without severe physical limitations, that is, an electronic dictionary, should provide all the inflected forms of a word, or at least that subset that suffices to establish for any inflectional form of a lexeme its exact morphological form (if any), and its phonetic form.

The language user may come across new words that (s)he has not seen or heard before, or may want to make a new word. For both purposes it is useful to include information on productive morphology in the dictionary. This is possible by making an entry for each productive affix. For unproductive affixes, on the other hand, it suffices to list all the existing words with that affix. In the entry for a productive affix, we specify the syntactic category of the base words to which the affix can be attached, and the meaning contribution of that affix to a complex word. Note that this meaning may co-vary with the word class of the base word.

As all morphologists know, it is not so easy to make a neat division between productive and unproductive affixes: some are semi-productive, that is, lead only occasionally to new formations. The best practical solution here is to also include affixes with a relatively low degree of productivity in the dictionary, because the language user may occasionally come across new words with such affixes.

In addition to productive affixes, there is also a large class of productive affixoids, morphemes that sometimes also exist as independent words, but always have a specific meaning when used in a complex word. For instance, the Dutch word *vrij* 'free' can be used as an affixoid, in combination with a noun, with the 'free from', as in Dutch *suikervrij* 'lit. free from sugar, sugarless'. The English morpheme *-free* behaves exactly the same way. Similarly, the Dutch word *oud* ‘ old’ can mean ‘ former, ex-‘ when part of a complex words, as in *oud-burgeneester* ‘ ex-mayor’. Clearly, such affixoids, which have arisen through grammaticalization of lexical words, require a lexical entry of their own. This also applies to the many neoclassical prefixoids that are used these days, such as *bio-, eco-, euro-*, borrowed morphemes that may correspond to words in the language of origin, but only occur as part of complex words in the borrowing language.

As pointed out above, existing compounds deserve at least being enumerated, in the lexical entry for the word that is the head, without further information on their formal and semantic properties. This is a good option for those compounds whose meaning is fully predictable on the basis of the meaning of the constituent words and the (conceptual and encyclopaedic) knowledge of the language user. When the compound has an unpredictable meaning, it should be given its own entry, however. Another reason for giving a fully regular compound its own entry is that there is more than one feasible interpretation, but only one of these is the conventional interpretation. For instance, the Dutch compound noun *waterbed* 'water bed' is normally used for designating mattresses filled with water, although it could also have been used for designating beds with which one can float on the water. Actually, this latter interpretation is still possible because productive word formation processes such a compounding are not absolutely blocked by existing lexical items. However, the dictionary should provide information about the conventional interpretation so that the language user realizes that the use of that word with another meaning might have a specific effect.

Regular compounds also have to be listed if they have an unpredictable linking element between the two constituents, as is the case in Dutch where *-s*, *-e* and *-en* may appear as linking elements. The choice of a linking morpheme is basically based on analogy to existing compounds

(Krott 2001). Listing of otherwise fully regular compounds is therefore necessary for a correct choice of the linking element.

Certain types of compounds stand in competition with phrases that are also used for designating categories. For example, in Dutch a number of types of cabbage are distinguished; some kinds are referred to by an Adjective-Noun compound, other kinds by an Adjective-Noun phrase:

AN compound: zuurkool 'sauerkraut '; spitskool 'oxheart / comical cabbage'
AN phrase: Chinese kool 'Chinese cabbage'; groene, witte, rode kool 'green, white, red cabbage',

We know for certain that, for instance, *rode kool* is a phrase because the adjective *rood* is inflected, which is impossible within a compound. Yet, such AN phrases are conventional lexical units that are functionally completely identical to compounds. Note also that the adjective in such AN phrases cannot be modified: a phrase such as *een heel rode kool* 'a very red cabbage' is no longer the name for a kind of cabbage, but a description of a particular cabbage. Therefore, the established AN phrases of this kind should be given in the entry for the head noun, so that the language user has clear information on the conventional labels in a particular domain, in this example the domain of cabbage.

The existence of such phrases has a blocking effect on the coinage of compounds: the formation of the compound *roodkool* is blocked by the existence of *rode kool*. This underscores the lexical status of such AN phrases, since competition with one winner (blocking) is characteristic for lexical units (cf. Jackendoff 2002 and Booij 2002 for detailed discussion of such lexical phrasal expressions).

4. Syntactic information and idiomatic patterns

A dictionary should specify which requirements a word imposes on its syntactic environment. Traditionally, this information is expressed by means of subcategorization features that indicate in which syntactic contexts a word can or must appear. Alternatively, one may use labels such as 'intransitive' and 'transitive' for verbs, and ‘count nouns’ versus ‘mass nouns’ for nouns. We should avoid, however, a too static interpretation of subcategorizational properties of words, since the grammar of a language provides means to change syntactic subcategory. For instance, the addition of a resultative predicate to a verb may change an intransitive verb into a transitive one, or may change the Aktionsart (type of event) of a verb. The Dutch verb *lopen* 'to walk' is an intransitive verb, but in combination with an adjective it is transitive, as in *Indriaas loopt zijn schoenen scheef* 'lit. Indriaas walks his shoes lopsided'. Conversely, transitive verbs can be used intransitively in the middle verb construction, as illustrated by the English sentence *These books sell well*. Moreover, there is a strong dependency of the syntactic valency of a word on its semantic interpretation. Therefore, the dictionary user should be made aware of the nature of such subcategorizational properties.

It is common wisdom that a dictionary should contain the existing, ithat is, established words of a language (however we determine exactly when a word exists), and all idiomatic word combinations. The notion 'idiomatic' should be understood here in a broad sense: not only word combinations of which the meaning is not fully compositional, but also word combinations that function as established, conventional units without having a non-compositional meaning. The term 'collocation' can be used for this more generalized interpretation of the notion 'idiom' (Everaert 1993). For instance, the fully transparent Dutch phrase *peper en zout* 'salt and pepper' has a fixed order for its constituent nouns, which is the opposite of the English order. Another example is the use of light verbs in combination with a noun, as in Dutch *een belofte doen* 'to make a promise, to promise'. The meaning of this phrase is transparent and compositional, yet one has to know that this is a conventional expression for the concept of promising.

A strong influx of multi-word expressions into the lexicon is caused by the phenomenon of grammaticalization (Hopper and Traugott 1993), the change of lexical morphemes into grammatical ones. This kind of change could already be seen above in the affixoids *-vrij* and *-free* which are halfway between lexical items and grammatical morphemes (affixes in that case). Many Dutch PPs function synchronically as prepositions, for instance *in verband met* 'in connection with, because of', and *met het oog op* 'lit. with the eye to, because of'. The NP *een paar* 'lit. a pair' functions as the quantifier 'some', witness the selection of a plural noun as in *een paar appels* 'some apples': the original head of this NP, *paar*, is singular, and yet we require the plural form of the noun *appels*.

Moreover, the number of apples is not necessarily 2 unlike what a literal interpretation of *paar* would imply. In sum, grammaticalized multi-word sequences must also be included in the dictionary.

It is important to realize that there are also syntactic units that are only partially idiomatic. Such idioms may be called idiomatic patterns or constructional idioms because the relevant set of expressions can be extended. A well-known example from Dutch is the construction exemplified by *een schat van een kind* 'lit. a sweatheart of a child, a sweet child'. In this construction, the noun of the complement functions semantically as the head noun, and the formal head noun as a modifier. This pattern *een N van een N* can also be extended to other nouns. Therefore, such patterns should be specified in a dictionary if the dictionary is conceived of as the storage house of all non-predictable information.

Some of these constructional idioms function as analytic lexical expressions, and there will therefore be no doubt that they must be dealt with in the dictionary. For instance, Dutch has a productive class of particle verbs (or separable complex verbs) of the type *door* + V, for instance *dooreten* 'to go on eating' and *doorzeuren* 'to go on nagging'. Such particle verbs have phrasal status because in Dutch main clauses the finite form of the verb appears in second position, whereas the particle appears clause-finally. Therefore, they cannot be seen as one word, because parts of words cannot be moved (the principle of Lexical Integrity). This class of constructions can be extended, and such verbs preceded by *door* 'through' have the systematic meaning 'to go on V-ing'. These patterns are idiomatic because the specific meaning of the construction is not fully derivable compositionally from the constituent words in isolation: *door* only has this specific meaning 'to go on' in combination with a verb. Hence, we should also create an entry for the word *door* used as a particle (it is also used as an adposition), and specify that, in combination with a verb, it expresses 'to go on with'. Thus, this specific entry for *door* will account for an idiomatic syntactic construction with a lexical function: these particles do what in other languages aspectual prefixes perform: the creation of verbs that express a specific kind of event (a specific Aktionsart) (cf. Booij 2002b). The use of such particles (also called preverbs) is a feature of many languages, also outside the Indo-European language family.

The example of *door* + V shows that productive syntactic patterns that create analytic lexical units should be specified in the dictionary just like productive word formation patterns. This pattern is not restricted to words that also function as adpositions. For instance, the Dutch adjective *open* 'open' also combines productively with verbs into an analytic lexical unit, functionally identical to verbal compounds, but formally a separable multi-wordunit. The unitary nature of such expressions is manifest in its syntactic behaviour. Compare, for instance, the established lexical unit *open maken* to the sequence *rood verven* 'to piant red'; the latter does not function as a unit in the progressive construction *aan het V*, unlike the former:

Jan is een fles aan het open maken / * Jan is een fles open aan het maken
John is a bottle at the open make-INF / John is a bottle open at the make-INF
'John is opening a bottle'

*Jan is een fles aan het rood verven / Jan is een fles rood aan het verven
John is a bottle at the red paint-INF / John is a bottle red at the paint-INF
'John is painting a bottle red'

Again, such analytic lexical units need be listed in the dictionary, and in addition, we need an entry for the adjective *open* that specifies its use as part of an analytic lexical expression, because this use of *open* is productive. Dictionaries of Dutch do list the existing cases of such multi-word lexical expressions, but the productive aspect of such patterns should also be accounted for.

The upshot of this section is that the syntax enters the dictionary, not only because of the existence of collocations, but also because certain kinds of phrases function as analytic lexical expressions, and form an alternative to expression of information by means of one grammatical word.

References

Booij, G. E. 1995. *The phonology of Dutch*. Oxford: Oxford University Press.

Booij, G. E. 2002a. *The morphology of Dutch*. Oxford: Oxford University Press.

Booij, G. E. 2002b. Constructional idioms, morphology, and the Dutch lexicon. *Journal of Germanic Linguistics* 14.4.

Chomsky, N. and M. Halle 1968. *The sound pattern of English*. New York: Harper and Row.

Everaert, M. 1993. Vaste verbindingen in woordenboeken. *Spektator* 22, 3-27.

Hopper. P. and E. Traugott 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Jackendoff, R. S. 2002. *Foundations of language*. Oxford: Oxford University Press.

Krott, A. 2001. *Analogy in morphology. The selection of linking elements in Dutch compounds*. Nijmegen: MPI (doctoral dissertation, Univ. of Nijmegen).

List of entries for the index

affixoid

allomorphy

collocation

compounding

constructional idiom

derivation

established word

existing words

grammaticalization

hyphenation

idiomatic pattern

lexical diffusion

lexical integrity

phonetic form

prefixoid

preverb

spoken language

stress

syllabification

word formation

underlying form

written language