For an informative and coherent classification of common nouns

Lucia M. Tovena
Universite de Lille
BP 149
F-59653 Villeneuve d'Ascq
tovena@univ-lille3.fr

Abstract

This paper deals with the interaction between linguistic data, as apprehended by a linguist, lexicographical description and computational lexicology. It concentrates on one class of words, that of common nouns, and looks at a case where a subset of mass nouns behave like count instead of siding with their fellow mass. This case raises the question of how to provide an informative and coherent classification for common nouns. The relevance of a proper classification of nouns is approached from the point of view of the characterisation of its combinatorics, an aspect relevant for all dictionaries especially in their function of language knowledge repositories for a learner, and from the point of view of the ontology behind the classification, an aspect relevant primarily for the development of sound lexical knowledge bases of the wordnet type.

Introduction

In this paper, we discuss the interaction between linguistic data, as apprehended by a linguist, lexicographical description and computational lexicology. We focus on one class of words, that of common nouns, and make a case for English and French.

Information on the combinatorics of words comes (primarily) in two forms in the definition of a lexical entry in a dictionary. First, essential information is conveyed via the grammatical characterisation[1], e.g. assigning a syntactic category, e.g. noun, and possibly adding features specific to the category, e.g. number for nouns and subcategorisation for verbs. Second, more information is provided in the form of multi-word expressions of various nature---e.g. idioms or collocations but also clauses exemplifying different instantiations of argument positions for a verb---and possibly examples that show types of variation of a given form, e.g. lexical variation in idioms such as *turn/tighten the screw on someone.*

In the following, we look at the issue of accounting for the distribution of nouns with respect to the first form of information, and its repercussions on the ontology underlying the classification. More specifically, we look at how to characterise a subset of mass nouns that behave like count instead of siding with their fellow mass, and we argue that a proper treatment of the denotation of common nouns can provide information translatable as intuitive and communicatively clear features that a lexicographer can use to capture restrictions on the lexical combinatorics of the entries, and that the computational lexicographers and computer scientists can exploit to build a sound lexical knowledge bases of the wordnet type.

Lines of classification

Common nouns form a complex group of linguistic entities on which several lines of characterisation have been applied in the literature. In a short and sketchy recall, we can mention four of them, all relevant for constraining the Det+N combination. Morphological number is a first line. It partitions common nouns into items with i) singular and plural forms, e.g. *house/houses*, ii) singular only, e.g. *relevance*, or iii) plural only, e.g. *scissors*. Number specification is given overtly in dictionaries only when the word has just one number form. Then, gender provides a line of characterisation which is particularly relevant for describing languages where gender is syntactically marked independently of its semantic import. For instance, the French dictionaries Le Grand Robert [LGR] and Trésor de la langue française [TLF] specify that *table* (table) is feminine and *livre* (book) is masculine. On the contrary, the English dictionaries Oxford Advanced Learner's Dictionary [OALD] and Longman Dictionary of Contemporary English [LDOCE] provide no gender information even for words such as *wife* whose gender is semantically relevant, e.g. the anaphoric reference via an overt pronoun. The learner is expected to guess it from her world knowledge.

Next, nouns can be partitioned into countable and mass. Most countable nouns are words for discrete entities that can be counted, like *apple* and *books*, and mass are usually words for entities that are thought of as a quantity or a substance, like *sand*. However, as just seen for gender, there isn't always a consistent correspondence between the linguistic 'mass/count' distinction and the more physical and philosophical sortal distinction, as one would expect if the former were established on the latter, hence the learner cannot infer the language specific value from world knowledge. This line is consistently exploited in English dictionaries but does not seem to belong to French lexicographic tradition, although in both languages it is relevant, e.g. for describing Det+N combinations. Information may come implicitly via examples, e.g. if the entry of countables contains the *un N* (a N) expression. The Dictionnaire du français langue étrangère Larousse [DFLE], that names the distinction overtly in its grammatical notes, is a noticeable exception.

The last line considered here partitions common nouns into concrete and abstract nouns. Most concrete nouns are words for physical entities, like *apple* and *milk*. Abstract nouns are usually words for processes, qualities and the like. This line is not clearly defined in theories constructed to explain the systems underlying a language, and mainly ignored in lexicography. In its general formulation, it may be problematic because it cuts across the countable/mass distinction. Members of the abstract group can be countable with singular and plural form, e.g. event names such as *destruction*, or mass with a singular form only, e.g. qualities such as *patience*. Indeed, all these lines can interact in several ways. In many languages countable nouns have singular and plural forms, whereas mass nouns have singular forms, but this is not a general rule. Often the use of a given line is related to its 'local' discriminative power, e.g Chinese has mass nouns only, which diminishes the interest of the third line for this language, but the choice may also have didactic purposes.

However, the mass/countable distinction is essential for understanding the functioning of nouns, and constantly referred to when discussing the semantic structure assigned to their denotation. The default domain of countable nouns is made up of individuals or sets thereof; for uncountable nouns, it is made up of parts [Krifka 1987], [Landman 1991]. On the contrary, the concrete/abstract distinction has not been used consistently and continuously in the linguistic tradition, but it relevance with respect to denotational issues has been recently argued for in formal terms in [Tovena 2001; 2002]. Several 'potential irregularities' in the distribution of singular existential determiners have received a uniform treatment thanks to a finer classification of nouns.

Why denotational issues are important to the learner

Surely, the mass/count distinction is a common feature of English and learners encounter it at a very early stage. The contrast in (1), discussed in any course for beginners, can be described by saying that the indefinite article *a* combines only with singular countable nouns. On the contrary, French lexicography lack a tradition of learner's dictionaries, and, as said above, traditional lexicography ignores the mass/count distinction, although it is equally important in understanding the functioning of the language, cf. the contrast in (2).

(1) a. a book
b.* a sand
c.* a books
(2) a. un livre (a book)
b.* un sable (a sand)

The point we are trying to make here is not in favour of a purported normalisation of lexicographic style across Europe, nor do we 'rediscover' the relevance of the count/mass distinction, rather we argue that, in both languages under discussion, there are cases where an even finer classification is actually needed in order to develop a learner's dictionary of real value. But before we come to that, a side remark on the fact that one could try to diminish the relevance of the mass/count distinction by recalling that English has a certain number of unrestricted determiners, i.e. determiners that combine with nouns of any number or type, e.g. the, *no, some, any*. Hence, the count/mass distinction is not always exploited in characterising Det+N combinations. In French, the sharp contrast in (2) might evade the attention of speakers and scholars because the grammaticality of (2b) can be 'improved' by adding a modifier to the noun, cf. (3a). However, modification is a strategy that gives access only to a taxonomic reading in case of mass nouns---i.e. the reading 'types of N'---while no such restriction applies to count nouns, cf. (3b). Thus, the grammaticality of (3a) does not diminish the relevance of the count/mass distinction. Furthermore, unrestricted determiners are not common in French.

(3) a. aucun sable fin (no fine sand) ? type reading
b. aucun livre fin (no thin book) ? instance or type readings

Evidence supporting the need for an even finer distinction comes from contrasts such as in (4), but there are several cases where determiners split the group of mass nouns in a systematic way. Such contrast can be replicated with other determiners, e.g. *le mondre* (the least) and *un certain* (a certain), and is present in English too, cf. (5).

(4) a. aucun livre (no book)
b.* aucun sable (no sand)
c. aucun courage (no courage)
(5) a. I have every confidence in him as a doctor
b.* I have every water in these pools

The contrasts in (4) and (5) can be tackled by referring to the distinction between traditional mass nouns and intensive quantities (IQ) [Van de Velde 1996], a kind of abstract mass nouns. IQs stand out for their possibility of undergoing continuous increase or contraction without a corresponding extension in space or time. This characterisation is communicatively and intellectually satisfactory. It underscores the interpretation as 'amount of an immaterial entity' intuitively prominent for these nouns, cf. (4), and is theoretically grounded [Tovena 2001; 2002]. The two types of mass nouns go different ways also in exclamatory and interrogative contexts, e.g. (6c) has only a rhetorical interpretation, if any.

(6) a. Quel livre veux-tu lire? (which book do you want to read?)
b. Quel vin veux-tu boire? (which wine do you want to drink?)
c.?? Quel courage a-t-il eu? (he didn't show any courage, did he?)

IQs behave like count nouns when looked at from the point of view of certain determiners, i.e. they seem to have a discretised denotational domain. But the units are not standard individuals, as they are not visible when looked at from other points of view, cf. (7).

(7) J'ai trois * courages/ ok livres (I have three courages/books)

IQ nouns are somewhere in between countable and mass. Tovena [2001] captures this situation by postulating three levels of discretisation for the denotational domain of a noun, namely via strongly or weakly discrete units, besides the case of no units at all. Strong units correspond to traditional atoms, hence are present in the domain of countable nouns only. They are the basis for telling to which group a noun belongs. The fact that IQ nouns belong to the mass group is derived from their nature as continuous (non-atomic) entities. At the same time, their 'non-standard' distribution in contexts of quantification seen in (4)-(5) is accommodated using information on weakly discrete units..

Why denotational issues are important to traditional and computational lexicographers

We endorse the idea that the domain of IQs contains weakly discrete units. The default way of discretisation for countables is by individuals. For the rest of mass nouns, the default strategy is by parts, defined with respect to units of measure, and by species only as a secondary option that applies to rescue a phrase, i.e. the taxonomic reading.

The distinction between strong and weak atoms can be translated into a form usable by a traditional lexicographer by exploiting the non-referential properties of units constituting amounts of IQs. Thus, examples such as (4) and (6) can be easily transformed into test grids for the classification of nouns, cf. [Tovena 2002] for more cases. The results can be expressed expanding the set of sortal features. This solution, which amounts to letting nouns subcategorise for determiners, like verbs for complements, is in line with proposals found in the literature [Pollard & Sag 1987].

Such distinction can be cashed in by computational lexicographers in terms of the distinction between the properties of *unity* and *identity* argued for in work on ontologies and information systems [Guarino & Welty 2000]. The notion of identity is intuitively linked with the issue of how complete is the description of an entity. The notion of unity is closely tied with that of parthood. Guarino and Welty show how these notions complement each other under the general notion of individuality, which is seen as the sum of the two. A countable noun can be characterised as naming an entity perceived as carrying both identity and unity properties (features). So, this description applies to *apple* but works also for a complex nominal expression like *piece of bread*, at least when 'piece' is intended as an undetached self-connected part of something. This captures the fact that countable nouns and noun phrases formed by a classifier---e.g. expressions such as *a piece* or *a slice*---and a mass noun go together in terms of denotation and syntactic functioning, cf. (8). They can be identified in a lexical database by the two features `+identity' and `+unity', where identity can be made more specific in terms of a given property.

(8) a. a/every/three book(s)
    b. a/every/three piece(s) of bread

Traditional mass, such as *bread*, carry no unity feature, as their parts can be arbitrarily scattered. They do carry identity features, based on the mereological extensionality of food, i.e. two amounts of food are the same if they have the same parts [Guarino and Welty 2000]. These nouns are identified by the combination `+identity –unity'

We propose a general way of characterising the notion of mass noun, based on the negativity of one of these two features, instead of taking as a crucial factor the absence of unity. This proposal is supported by the consideration that the absence of one of the two features results in the absence of individuality typical of all mass nouns. This move allows us the possibility of generalising over mass nouns, but at the same time it allows for two distinct subcases, as either unity or identity is present. Thus, mass nouns in general can be identified in the database by the two combinations of features `+identity –unity' or '-identity +unity'. Traditional mass nouns are characterised by the first combination. IQs' weakly discretised domain is composed of units that partition the domain, hence it is composed of entities that carry a unity feature but do not qualify as individuals, thus do not have an identity feature. So, IQs are identified by the combination '-identity + unity'. Then, the use of Guarino and Welty's [2000] features makes it possible to impose some constraints on the IS-A relation, as it affects the possibilities of subsumption among objects. This results in a cleaner taxonomic organisation.

Conclusion

This paper has drawn the attention on the behaviour of a subset of mass nouns, dubbed IQs, that side with singular countable nouns, rather than plural and mass at large, in certain quantificational contexts. We have adopted Tovena's [2001] proposal---that between the non atomic domain of masses and the atomised domain of count nouns there is a third type made of weakly discretised units---as the basis for an informative lexicographical distinction. The finer grained classification, based on grammatical tests, and possibly enriched with labels from a descriptive classification [Flaux & Van de Velde 2000], gives room for describing different grammatical statuses for given Det+N combinations. Finally, the use Guarino and Welty's [2000] features results in a clean taxonomic organisation.

Endnotes

[1] One reviewer objected that the term 'combinatorics' cannot be used to describe the behaviour of words resulting from categorial or other grammatical properties. Such a position is questionable for at least two reasons. First, syntactic category is at the basis of many cases of lexical combinatorics, e.g. in subcategorisation, VP idioms, collocational adjectives, etc. Second, the dichotomy between 'grammatical' and 'lexical' results from a choice. It has disappeared in lexicalist theories of the grammar, e.g. HPSG [Pollard & Sag 1987]. The objection that 'lexically restricted co-occurrence' does not mean 'semantically restricted' for a lexicologist is also dubious; verb subcategorisation provides a good counterexample again. In the case in hand, giving up a fine semantic characterisation of means to treat data like (4) as idiosyncratic and to deprive the learner of a valid criterion.

References

[DFLE] *Dictionnaire du français langue étrangère Larousse*, 1978

[Flaux & Van de Velde 2000] Flaux N. and D. Van de Velde, 2000. *Les noms en français.* Ophrys, Paris.

[Guarino & Welty 2000] Guarino, N. and Ch. Welty, 2000. Identity, Unity, and Individuality: Towars a Formal Toolkit for Ontological Analysis. *Proceedings of ECAI.* IOS Press.

[Krifka 1987] Krifka, M., 1987. Nominal reference and temporal constitution: Towards a semantics of quantity. *Forschungstelle für natürlichsprachliche Systeme, Bericht 17.*

[Landman 1991] Landman, F., 1991. *Structures for Semantics.* Kluwer Academic Press, The Netherlands.

[LGR] *Le Grand Robert de la langue française.* second edition 1985

[LDOCE] *Longman Dictionary of Contemporary English.* third edition 1987

[OALD] *Oxford Advanced Learner's Dictionary.* sixth edition 2000

[Pollard & Sag 1987] Pollard C. and I. Sag, 1987. *Information based syntax and semantics.* CSLI.

[Tovena 2001] Tovena, L. M., 2001. Between mass and count. *Proceedings of the Western Coast Conference on Formal Linguistics.* pp. 565-578. Sommerville MA: Cascadilla Press

[Tovena 2002] Tovena, L. M., 2002. Determiners and weakly discretised domains. *Going Romance 2001: selected papers.* Jonh Benjamins, The Netherlands. (in press)

[TLF] *Trésor de la langue française.* 1979

[Van de Velde 1996] Van de Velde, D., 1996. *Le spectre nominal.* Peeters, Paris.