Toni BADIA & Roser SAURÍ, IULA, Universitat Pompeu Fabra

Polysemy and underspecification of *bottle* and related nouns

Abstract

In this paper we discuss containers and other general nouns, and develop a proposal for representing them in a structured lexicon. We adopt a typed feature structure formalism and show that even in more cases than those mentioned in the literature an underspecification analysis is appropriate. This contributes to the simplification of the lexicon, postulating less lexical rules and avoiding a lot of redundancy. Our main data come from Catalan, but the results are applicable to many other languages (including English). The paper is organised as follows. In section 1 we present the Catalan data. In section 2 we discuss some of the previous proposals. Section 3 is devoted to develop our treatment, which is implemented in LKB.[1] The main conclusions are given in section 4.

Keywords: computational lexicography, polysemy, structure of the lexicon, nominal complements

1. General nouns in Catalan

There is a relatively closed set of general nouns which very often belong to the specifier of an NP. They can be grouped as expressing collectivity (1a), containers (1b), measure (1c), or partition (specific (1d), or general (1e)):

(1) a. ram ('bouquet'), grup ('group')
b. ampolla ('bottle'), vas ('glass')
c. litre
d. llesca ('slice (of bread)'), full ('piece (of paper)')
e. mica ('bit'), tros ('portion')

These nouns contribute to the semantic operations of individuation, quantification and specification which are generally performed by determiners. In (2a), for example, *una ampolla* delimits a specific quantity of wine, indeed in a much more precise way than *molt de* and *poc* do.

(2) a. una ampolla de vi ('a bottle of wine')
b. molt de vi ('much wine')
c. poc vi ('little wine')

From a syntactic point of view they are, together with the appropriate determiners and with the preposition *de*, in complementary distribution with other determiners; for example, in (2) *una ampolla de* alternates with *molt de* and *poc*.

There have been some proposals to treat constructions like *una mica de* ('a bit of') as complex determiners. However, there are strong reasons for adopting a nominal category analysis, i.e., that these nouns head an NP to which a PP with *de* ('of') is attached. In topicalised or dislocated constructions the preposition *de* ('of') appears with the dislocated (topicalised) noun, thus breaking the supposed complex determiner:

(3) De roses, en vull un ram ('of roses, CLI I want a bouquet')

In these constructions the preposition *de* introduces the complement to the noun. In Catalan (and in other languages) we find some evidence in favour of a strict subcategorisational relation: the complement of the general noun can only be elided (or topicalised) if the partitive clitic *en* is present in the verbal environment:

(4) a. De llet, en vull un litre ('Of milk, CLI I want a litre')
b. * De llet, vull un litre ('Of milk, I want a litre')

(5) a. Quanta llet vols? En vull un litre ('How much do you want? CLI I want a litre')
b. * Vull un litre ('I want a litre')

Note that the partitive clitic *en* also appears in the elision or topicalisation of nouns specified by a quantifier:

(6) a. Vull dues/algunes pomes ('I want two/some apples')
b. De pomes, en vull dues ('Of apples, CLI I want two')

There are some cases however where the complement does not overtly appear. Firstly, some specific partitives and collectives can be construed without a complement but with an implied semantics (7a)–(7b). And secondly containers can be used in such a way when denoting a container (and not the containee) (7c) and so are measure nouns, when interpreted as a purely abstract quantity (and not as a specific quantity (7d)). Therefore the nouns belonging to these two latter groups are semantically ambiguous. When interpreted in one sense a complement is necessarily overtly realised; and when interpreted in the other it is not.

(7) a. Vull dues llesques ('I want two slices (of bread)')
b. Vull un ram ('I want a bouquet (of flowers)')
c. Vull un vas ('I want a glass')
d. Pesa tres quilos ('It weights three kilos')

Containers can denote either a physical object (container) (8b) or the quantity of mass contained (containee) (8a), depending on the context:

(8) a. He begut dos vasos de vi ('I've drunk two glasses of wine')
b. He trencat dos vasos de vi ('I've broken two glasses of wine')
c. Hi ha dos vasos de vi ('There are two glasses of wine')

In elision and topicalisation contexts these two semantic interpretations have a different behaviour: again the presence of the clitic *en* is implied in the former, but not in the latter:

(9) a. De vi, n'he begut dos vasos ('Of wine, CLI I've drunk two glasses)
b. De vi, he trencat dos vasos, i d'aigua, quatre ('Of wine, I've broken two glasses, and of water, four')

2. Previous proposals

In this section we describe previous treatments of general nouns in typed formalisms. For reasons of space we only discuss a small, but representative, sample of them. We first present the basic elements of each proposal, and then discuss their ability to the account for the linguistic facts.

Copestake (1992) (developing an original proposal by Ostler & Atkins (1991)) offers a treatment of the sense alternations of these nouns in the LKB framework. Ostler and Atkins had highlighted a number of regularities in the sense alternations of nouns, such as animal–meat, animal–skin, tree–wood, tree–fruit, container–containee, and so on. The basic insight of Copestake's (1992) is that nominal denotations are classified according to their ontological character, and sense alternations are represented as links between different ontological types. The classification is implemented as a type system in LKB in which lexical semantic properties of nouns are structured along the lines of Pustejovsky's *Generative Lexicon* (1991; 1995). And the links which relate in a regular way the different types that correspond to the various senses of a noun are implemented as lexical rules. Thus the animal–meat sense alternation present in the use of a noun as *rabbit* is accounted for by a couple of types corresponding to a living animal and to a mass of meat respectively. These two types are then related by a specialisation of the general grinding rule which takes an individual and provides a mass.

This proposal allows for the treatment of lexical sense variation in a parallel way to morphological operations. This is particularly appropriate since in some languages there are sense alternations which imply a morphological change, as in the Spanish tree–fruit alternation; in these cases the very same rule that expresses the semantic variation can express the morphological operation. An improved proposal on the same line is Copestake & Briscoe (1996); they include the contextual information to deal with some kinds of sense alternations; however they still maintain lexical rules for some other regular alternations, such as containers.

A different proposal to treat sense alternations in LKB is Climent (1996). He deals with the individual–mass alternation by determining a small set of modes of reference, which are distinguished from one another by a pair of distinctive features: discretivity and multiplicity. A simple type corresponds to each mode of reference and types are linked by four lexical rules: individuation, pluralisation, cumulation and generalisation. The lexical characterisation of nouns provides them with a type (in which some of the elements of Pustejovsky's qualia structure are present); and lexical rules apply to them insofar as the types allow their application.

An interesting aspect of both these analyses is that they introduce a complex semantics for lexical signs (following more or less closely Pustejovsky (1995)). A full lexical semantic description is indeed needed to treat sense alternations as these in an adequate way, and the *Generative Lexicon* framework provides such a description. Whereas Climent (1996) attempts to explain only the mass–countable distinction with a single, very simple, mechanism, Copestake (1992) offers a thorough treatment of a large range of linguistic phenomena. As a result, Copestake's proposal is much more adequate from a descriptive point of view: it gives a precise account of a reasonably great number of well established linguistic sense alternations. The appropriatness of such a treatment to our view relies on at least three formal

devices: a detailed structuring of the lexical semantics of a nominal sign, a complex type system that build upon the semantic structure, and a set of very specific lexical rules.

The heavy use of lexical rules, however, results in some problems (some of them well known): as implemented in LKB they are unidirectional, their application cannot be controlled (i.e., lexical entries cannot restrict their application), and they cannot allow interaction with the context of use of the noun. The overall system, with sense alternations expressed with lexical rules, does not benefit from the descriptive power of the complex structure and typing; once types are complex enough to describe the differences in linguistic behaviour between, say, different containers, lexical rules are either too general, or lexically dependent. In the first case they overgenerate and in the second one they become too numerous to be adequately controlled by the lexicographer. In addition, sense alternations are very often contextually induced; and this relation between context of use and a particular sense obviously cannot be captured by lexical rules (which apply within the lexicon).

We thus try to provide a system with a complex semantic structuring and typing (very similar in spirit to Copestake's (1992)) in which the expressive power of the type system is used to determine the allowable sense alternation. The formal devices used are underspecification and type resolution (which in common provide an implementation to Pustejovsky's notion of type coercion).

3. Formal treatment

We thus attempt to account for the semantic ambiguity of container nouns on a par with the subcategorisation differences in the possible senses. We adopt the lexicalist approach of HPSG where it is possible to give a simultaneous treatment of syntax and semantics. However its semantics has to be enriched with the lexical semantics of Pustejovsky (1995) to cope with all sense distinctions that are needed. The appropriate type for common nouns results as follows:

(10)
$$\left[\text{LOC:}\left[\text{CAT:}\begin{bmatrix}\text{HEAD: } \textit{noun}\\ \text{COMPS: } \textit{list-of-synsem}\end{bmatrix}\right.\right.$$
$$\left.\left.\text{CONT:}\begin{bmatrix}\text{IND: } \boxed{2}\\ \text{RESTR: } \left\{\underset{psoa}{\begin{bmatrix}\text{REL: } \textit{rel}\\ \text{INST: } \boxed{2}\end{bmatrix}}\right\}\\ \text{QUALIA: } \underset{qualia}{\begin{bmatrix}\text{FORMAL: } \textit{formal}\\ \text{CONSTIT: } \textit{constit}\\ \text{TELIC: } \textit{telic}\end{bmatrix}}\end{bmatrix}\right]\right]$$

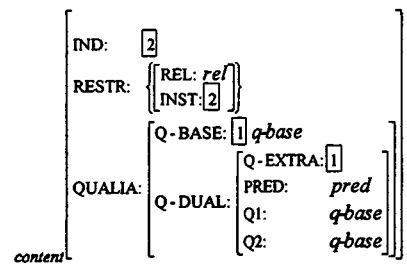
In standard HPSG the content structure of a nominal sign is of type *nominal-object*, and bears the INDEX (IND) and RESTRICTION (RESTR) attributes. The former acts as the referent marker, whereas the latter expresses the relation that the noun holds. Notice however that in our proposal the content structure has been enlarged with the QUALIA feature, which corresponds to the descriptive object used by Pustejovsky (1991; 1995) to express the semantic information of linguistic elements. As will be seen, in this treatment the RESTR structure provides the arguments to which, by means of coindexation, the QUALIA structure will assign the

properties. We thus assimilate the initial PSOA in the RESTR set to Pustejovsky's argument structure (ARGSTR).

In Copestake's approach, the different senses of a polysemic noun such as *bottle* are represented by means of a lexical rule relating the QUALIA structures of two different lexical entries. This introduces much redundancy in the lexicon. However such an undesirable consequence can be avoided by applying Pustejovsky's notion of dotted type (Pustejovsky, 1995:95). This notion originates in the *Lexical Conceptual Paradigm (lcp)* developed by Pustejovsky & Anick (1988) to refer to the lexical items that cluster multiple related senses. The notion of *lcp* allows to treat the different senses of the same word not as distinct, but as logical expressions of different aspects of a single meta-entry. The appropriate types for this specific entries are the dotted types. Thus, dotted types in the hierarchy express relations that generally hold between different senses of polysemic words, such as the relation observed in container nouns between the container and the containee. To this dotted object point the meta-entries of container nouns such as *bottle*. This obviously results in a remarkable reduction of lexical information.

In our computational framework (LKB), the dotted type notion appears to be easily implementable by means of type underspecification; that is, building up a type where the two senses are expressed, without resolving in favour of only one of them. To that purpose, some formal aspects must be considered. First of all, the appropriate level to express the dotted relation should be the QUALIA structure, since it is here where the lexical semantic information is captured. Notice that in our proposal the dotted relation is expressed by means of the whole qualia structure, differently than Pustejovsky's treatment where only the FORMAL attributes are related. Secondly, the formal device to express the polysemy will be a complex qualia structure (Q-DUAL). It is composed by a first attribute which expresses the relation between the two senses (PRED), and two other attributes representing the two possible senses, called QUALIA-1 (Q1) and QUALIA-2 (Q2). Thirdly, to select the appropriate sense in each context a fourth attribute in Q-DUAL, called Q-EXTRA, is added; the sense selection is done by coindexating Q-EXTRA with Q1 or Q2, depending on the sense appropriate in each context. Finally, the attribute Q-BASE allows the extraction of the selected sense out of Q-DUAL, by means of coindexation with Q-EXTRA. The resulting CONTENT structure for dotted types is shown below:

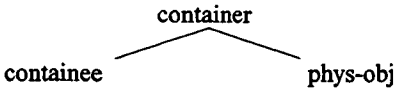
(11)



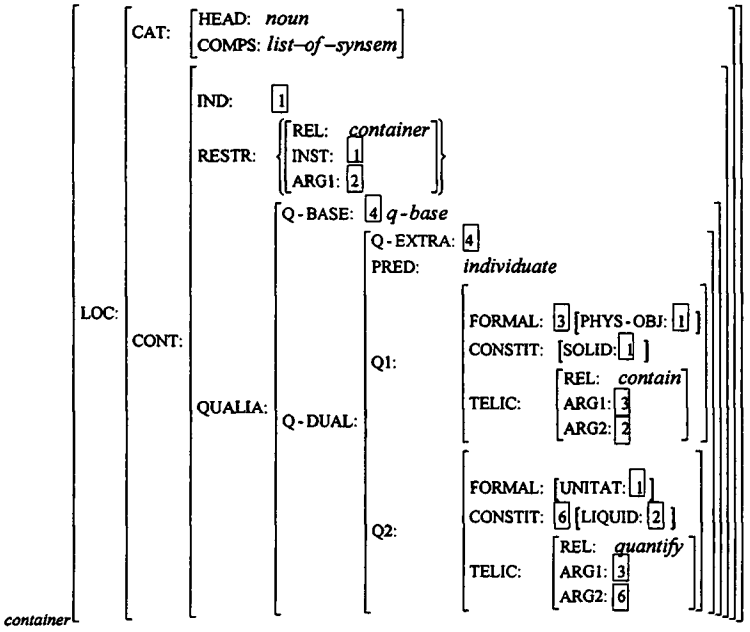
In LKB this is obtained by means of type resolution. Thus the dotted type must present two specific subtypes: a first one in which the Q1 information is captured via Q-EXTRA and a second one in which the selected information is Q2.

Let us now apply this treatment to container nouns. The partial hierarchy (12) is composed by their general type (13) and the two subtypes shown in (14) and (15). The first one represents the containee sense, and its SUBCAT list asks for a specific complement.[2] The second one represents the physical object sense, its SUBCAT presents an empty list as its appropriate value and the set of thematically bound adjuncts (Sanfilippo, 1998) restricts the modifiers implied by the qualia of the noun.[3] In what follows, such a mechanism is exemplified for a container noun like *bottle* when occurring in contexts similar to (8a) or (8b). Consider the partial type hierarchy developed for containers, and let the lexical entry for *bottle* point to the general type (13).

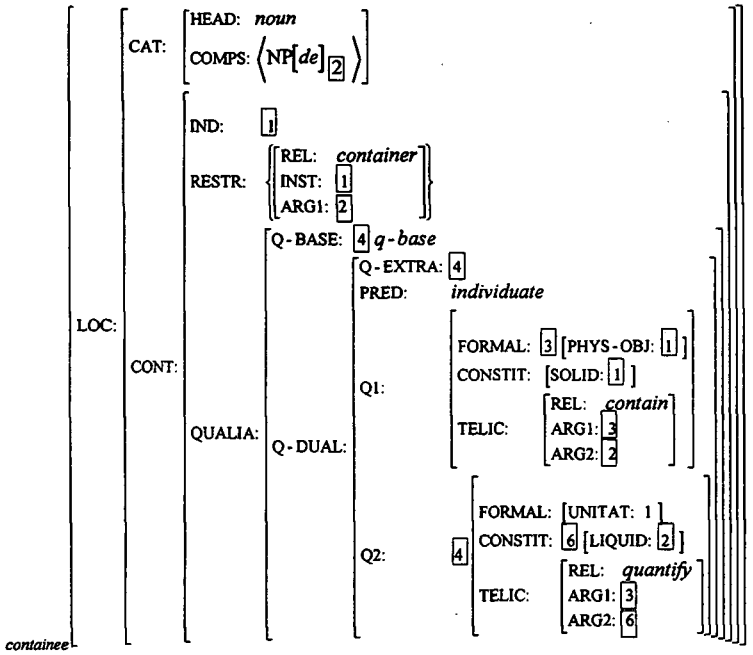
(12)



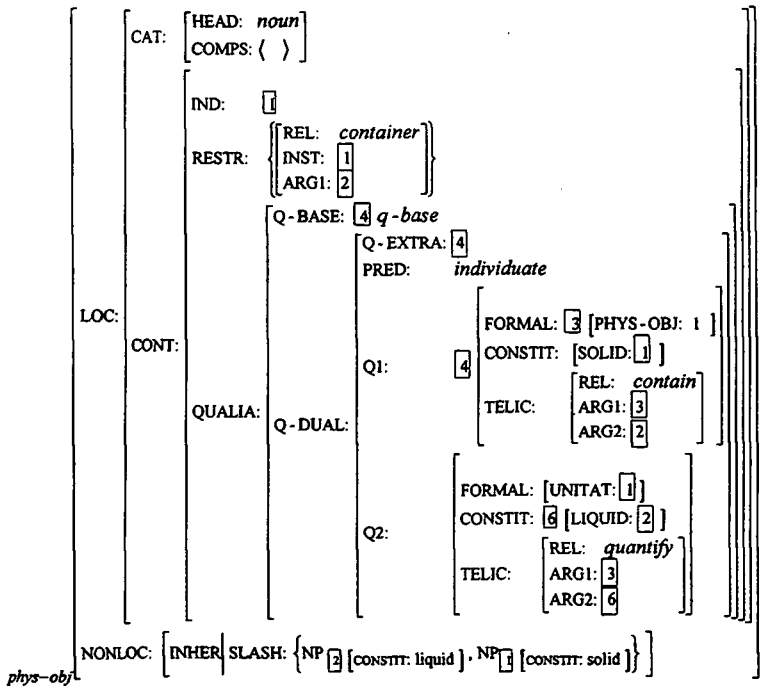
(13)



(14)


containee
LOC:
CAT:
HEAD: noun
COMPS: ⟨ NP[de] 2 ⟩
CONT:
IND: 1
RESTR: { REL: container, INST: 1, ARG1: 2 }
QUALIA:
Q-BASE: 4 q-base
Q-DUAL:
Q-EXTRA: 4
PRED: individuate
Q1:
FORMAL: 3 [PHYS-OBJ: 1]
CONSTIT: [SOLID: 1]
TELIC: [REL: contain, ARG1: 3, ARG2: 2]
Q2: 4
FORMAL: [UNITAT: 1]
CONSTIT: 6 [LIQUID: 2]
TELIC: [REL: quantify, ARG1: 3, ARG2: 6]

(15)


phys-obj
LOC:
CAT:
HEAD: noun
COMPS: ⟨ ⟩
CONT:
IND: 1
RESTR: { REL: container, INST: 1, ARG1: 2 }
QUALIA:
Q-BASE: 4 q-base
Q-DUAL:
Q-EXTRA: 4
PRED: individuate
Q1: 4
FORMAL: 3 [PHYS-OBJ: 1]
CONSTIT: [SOLID: 1]
TELIC: [REL: contain, ARG1: 3, ARG2: 2]
Q2:
FORMAL: [UNITAT: 1]
CONSTIT: 6 [LIQUID: 2]
TELIC: [REL: quantify, ARG1: 3, ARG2: 6]
NONLOC: [INHER | SLASH: {NP 2 [CONSTIT: liquid], NP 1 [CONSTIT: solid]}]

In the cases in which some contextual element asks for *bottle* as a physical object, the system will resolve the interpretation in favour of the type (15); whereas in the cases where a (quantity of) liquid sense of *bottle* is required, the type chosen will be the one in (14). The former operation applies in contexts similar to the one in (8b), since the verbal predicate *trencar* ('break') asks for a solid physical object complement. Cases similar to (8a), in which the verb *beure* ('drink') requires a liquid complement, resolves in favour of the second type resolution operation. A great achievement of this proposal is that in those contexts where the polysemous noun is not disambiguated (8c) the general type can be maintained without resolving into a particular subtype.

This treatment is also applicable to the nouns of the same class. Measure names which are also polysemic (the abstract measure and the specific quantity of an object or a mass) can be dealt with in an exactly parallel way to containers. Other non polysemic nouns can be represented by one of the specific subtypes shown above: *mica* ('bit') would point to a type similar to the containee sense, and *llesca* ('slice'), to the physical object type.

4. Conclusions

In this paper we have shown that to give account of the behaviour of the nouns introduced at section 2 it is needed a complex semantic characterisation of every word in the lines developed by Pustejovsky (1995). The use of both the dotted object notion and the type coercion mechanism (implemented in LKB by means of type underspecification and type resolution respectively) can adequately deal with the polysemous nature of some of these words, and determine the selection of the appropriate sense considering the context of use. This approach differs from the one proposed in Copestake (1992) in that not only the prototypical Pustejovsky's structuring of the semantic information is used, but also his generative mechanisms.

An interesting achievement of this approach is the capacity of including contextual information when building the semantic interpretation of a given construction. As has been seen, it supposes a strong mechanism to constraint the interpretation of ambiguous nouns and reduces the use of lexical rules.

5. Notes

[1] *Lexical Knowledge Base* (Copestake, 1992) is an implemented platform to represent lexical knowledge.

[2] Observe that the information expressed in the QUALIA structure (specifically in the PRED attribute of the TELIC structure) allows to nicely explain the similarity between the function of determiners and the function of this group of nouns.

[3] Note that this notion of thematically bound adjuncts implements in our system the default arguments of Pustejovsky's (1991; 1995).

6. References

Climent, S. (1996) Modes of reference. Representation and derivation from prototypical specifications, in *Procesamiento del Lenguaje Natural, 19. Actas del XII Congreso.* Sociedad Española para el Procesamiento del Lenguaje Natural, Barcelona.

Copestake, A. (1992) *The Representation of Lexical Semantic Information.* Doctoral dissertation, University of Sussex, Cognitive Science research paper CSRP 280.

Copestake, A. & T. Briscoe (1992) Lexical operations in a unification based framework, in J. Pustejovsky & S. Bergler (eds.) *Lexical Semantics and Knowledge Representation. Proceedings of the First SIGLEX Workshop.* Berkeley, CA, Springer-Verlag, Berlin, pp. 101–119.

Copestake, A. & T. Briscoe (1996) Semi-productive polysemy and sense extension, in *Proceedings of the Eighth European Summer School in Logic, Language and Information.* August 12–23, 1996, Prague.

Ostler, N. & B.T.S. Atkins (1992) Predictable meaning shift: some linguistic properties of lexical implication rules, in J. Pustejovsky & S. Bergler (ed.) *Lexical Semantics and Knowledge Representation. Proceedings of the First SIGLEX Workshop,* Berkeley, CA, Springer-Verlag, Berlin, pp. 87–100.

Pollard, C. & I. Sag (1987) *Information-based Syntax and Semantics. Vol. 1: Fundamentals.* CSLI Lecture Notes, 13. Stanford, Center for the Study of Language and Information.

Pollard, C. & I. Sag (1994) *Head-driven Phrase Structure Grammar.* Studies in Contemporary Linguistics. Stanford, CSLI and Chicago, University of Chicago Press.

Pustejovsky, J. (1991) The generative lexicon, in *Computational Linguistics,* 17 (4), 409–441.

Pustejovsky, J. (1995) *The Generative Lexicon.* MIT Press. Cambridge, MA.

Sanfilippo, A. (1998) Thematically bound adjuncts, in S. Balari & L. Dini (eds.) *HPSG in Romance.* CSLI Lecture Notes. Stanford, Center for the Study of Language and Information.