The hyperonym problem revisited: Conceptual and lexical hierarchies in language generation

Manfred Stede
Technical University of Berlin
Dept. of Computer Science
KIT Project Group
10587 Berlin/Germany
stede@cs.tu-berlin.de

Abstract

When a lexical item is selected in the language production process, it needs to be explained why none of its superordinates gets selected instead, since their applicability conditions are fulfilled all the same. This question has received much attention in cognitive modelling and not as much in other branches of NLG. This paper describes the various approaches taken, discusses the reasons why they are so different, and argues that production models using symbolic representations should make a distinction between conceptual and lexical hierarchies, which can be organized along fixed levels as studied in (some branches of) lexical semantics.

1 Introduction

Representations used in language processing owe much to the tradition of 'semantic networks', which nowadays have been successfully formalized and organized especially around one particular kind of link between nodes: the ISA-link, which connects entities to subordinate entities. This link is, by definition, the root of the so-called 'hyperonym[1] problem': When a speaker utters a word, she presumably needs to retrieve a lemma from her mental lexicon, and the 'applicability conditions' of the lemma automatically render the lemma's hyperonyms also applicable, thus raising the question how the choice among a set of more or less specific words is made.

In this paper, I briefly review approaches to the hyperonym problem in psycholinguistics, natural language generation, and lexical semantics. In doing that, I will refer to different branches of NLG according to their roots and main motivations. Generally acknowledged are the two poles of 'cognition-inspired' and 'engineering-inspired' language production: Cognition-inspired work (CI-NLG, for short) seeks to build models that replicate performance data and explain phenomena of human language production with the help of psychological experiments; engineering-inspired work (EI-NLG) seeks to build programs that provide linguistic output to some particular computer application. These goals are extremely different, and it seems that the gap between the respective methodologies will persist for quite some time. In between the two, however, I would situate a third category, which may be called 'linguistics-inspired'. For this branch, here abbreviated as LI-NLG, the primary motivation is neither in modelling human performance nor in efficiently performing a technical application; rather, LI-NLG seeks production models that replicate 'competence data', i.e. that account for observed linguistic regularities, without commiting to statements about the human production *process*.

Arguing that progress hinges on a better understanding of the structure of the mental vocabulary, which includes a clear picture of the nature of the ISA-link, I will sketch a framework of distinct (but related) conceptual and lexical hierarchies, which offers possibilities to account for at least some of the phenomena to be discussed.

2 The hyperonym problem

Following the psycholinguistics literature, the hyperonym problem is regarded as an aspect of *lemma retrieval*. Roelofs [1996, p. 308] describes a 'lemma' as a representation of the meaning and the syntactic properties of a word, and the task of lemma retrieval as a crucial step in the

[1] Alternatively called 'hypernym' in many publications; 'hyperonym' seems preferable, as the Greek root is 'hyper' (super) + 'onoma' (name).

process of grammatical encoding, where building of a phrasal, clausal, or sentential structure requires the syntactic information that lemmas contain.

Thus abstracting from the other steps of language production (formulation, articulation) as well as from possible influences of context, the task is confined to retrieve a lemma that corresponds to the conceptual specification that is represented in some adequate way. For the psycholinguist, the general problem is that of *convergence* from an under-specified conceptual representation to one word that the speaker utters. Levelt [1989, p. 201] characterizes the hyperonym problem:

> "There is one particularly nasty convergence problem that has not been solved by any theory of lexical access. I will call it the hyperonym problem [...]: When lemma A's meaning entails lemma B's meaning, B is a hyperonym of A. If A's conceptual conditions are met, then B's are necessarily also satisfied. Hence, if A is the correct lemma, B will (also) be retrieved."

The relation of hyperonymy is generally regarded as transitive: If A is a hyperonym of B, and B is a hyperonym of C, then A is a hyperonym of C. Following common practice, we call A a *direct* hyperonym of B, while it is only an indirect hyperonym of C. The same holds for the inverse relation, hyponymy.

For CI-NLG, which is concerned with finding models that resolve the convergence problem with the impressive speed displayed by human speakers, the hyperonym problem is important because it serves to put implemented models of spreading activation to the test. For EI-NLG, on the other hand, it can usually be ignored, as most of today's practical applications either do not require the production of a more general word (i.e., there is a one-to-one mapping from concept to word) or can rely on fairly simple mechanisms that avoid lexical repetitions by choosing from a fixed, pre-defined set of near-synonyms. For LI-NLG, the challenge of the hyperonym problem is to explain how a sentence can be paraphrased by others that replace a word by a hyperonym, and why speakers select from candidate hyperonyms in different situations of utterance. More concrete, given a conceptual specification (in a wide sense, including contextual parameters and communicative goals), the task is to find the best candidate from a set of valid paraphrases, here especially on the grounds of replacing content words with hyperonyms.

3 Psycholinguistic production models

Language production models developed in psycholinguistics are nowadays couched in neural network theory. Under debate are the computational properties of the networks, i.e., the modes of activation spreading, the existence of feedback, of inhibitory links, etc. The main methodological concern is to construct the models in such a way that they account for data gathered in human speech production experiments, often involving production errors, which can shed light on the underlying mechanisms.

A central point of content is the question whether the meaning of concepts and/or words is represented in a *decomposed* fashion or not. Here, the hyperonym problem is sometimes used as evidence by proponents of non-decompositional models. Roelofs [1996], for instance, argues that if a number of nodes representing semantic features are the basis for lexical access, in lemma retrieval it becomes extremely difficult to control the activation spread in such a way that only the most specific lexical unit that combines these features gets selected. Roelofs concludes that a non-decompositional model is to be favoured: When lemma retrieval starts with activation of the 'lexical concept' FATHER, rather than with the features MALE and PARENT, the output word will be *father*, without the danger of being outranked by a higher activation of *parent* (or *person*, or *entity*, presumably).

This line is continued in a recent comprehensive theory of speech production by Levelt, Roelofs, and Meyer [1999]. The focus of this theory is more on the side of articulation, but their approach to (non-) decomposition and hyperonyms follows the basic assumption just sketched. The model consists of three layers of nodes: A layer of concept nodes with labelled concept links, a layer of lemma nodes, and a layer of word form nodes that include morphological information. When a lexical concept is activated, the mechanism of activation spreading ensures that the directly connected lemma receives the highest activation, and not a lemma associated with a hyperonym of the lexical concept (which is connected by an ISA-link).

Working out the mechanics to ensure this behaviour is important for the implementation, but from the particular viewpoint of word choice, approaches of this kind are not very explanatory. Levelt et al. [1999, p. 4] state that "there is not the slightest evidence that speakers tend to produce hyperonyms of intended target words." But when lexical access *starts* with an appropriately activated lexical concept, the problem is effectively moved away, into the realm of conceptualization. The authors acknowledge the need for a component that establishes a 'perspective' by selecting a specific set of words, but have not incorporated such a component into their model. Thus, why and how the lexical concept receives its activation, and where the intention of using a word arises from, is not covered by the theory. For these questions, we have to turn to work in natural language generation.

4 Hyperonyms in NLG systems

In contrast to psycholinguistics-inspired work, the vast majority of natural language generation systems uses computations based on symbol manipulation, often connected with symbolic knowledge representation and reasoning techniques. In these systems, the hyperonym problem as one aspect of the general task of *lexical choice* arises only in systems that employ a sufficiently rich model of the lexicon and the concept–lexicon link, involving some sort of hierarchy information. As pointed out above, from an application-oriented perspective (i.e., in EI-NLG) it is often sufficient to work with rather limited mechanisms that largely eschew the lexical choice task.

The earliest and very influential device for performing lexical choice, Goldman's [1975] *discrimination net* hard-wires the sequence of choice points leading to a specific lexical item, which is in fact the general strategy taken in the majority of NLG systems: if you have a choice, then prefer the most specific term.

The most substantial criticism on the prefer-the-specific heuristic has been voiced in the work of Reiter [1991]. One of his examples is a system answering the question *Is Terry a woman?* Even if the system has the specific knowledge that Terry is a bachelor, the response *No, Terry is a bachelor* would not be appropriate here; the less specific *No, Terry is a man* is better since it does not prompt the hearer to draw any conclusions as to the particular relevance of Terry's marital status for the present conversation. Reiter's main point is to distinguish the *knowledge* a generation system has at its disposal from the *communicative goals* followed in producing an utterance. The latter are explicitly represented in his system as a list of attributes 'to communicate about an entity', which is a subset of the overall knowledge the system has of that entity. In the Terry-example, the goal is to inform the hearer that Terry has the attributes {Human, Age-status:adult, Sex:Male}.

In the KL-ONE [Brachman, Schmolze 1985]) style knowledge representation used by Reiter, concepts can be marked as 'basic-level' in the sense of [Rosch 1978]. Thus, on the taxonomic path *Tweety* (instance-of) *Robin – Bird – Vertebrate – Animal – Object*, the concept *Bird* is a basic-level one, which leads to a preference for using the corresponding lexical item when referring to some kind of bird (i.e., some concept or instance subsumed by it). Simultaneous to Rosch's work, Cruse [1977] (who in turn was building on earlier research by Roger Brown in the 1960s) had pointed out that the failure to use items of "inherently neutral specificity" (a notion that closely corresponds to the basic level) results in unwanted conversational implicatures — the hearer will surmise the existence of some reason why the neutral term *could not* be used in the specific situation of utterance.

But using the basic level is not mandatory, of course. Given a suitable context where attention is directed to particular attributes of entitities, a speaker moves to a more specific or sometimes to a more general level. Reiter's mechanism of to-communicate attributes tries to capture this: Covering these attributes with a suitable term can override the preference for the basic level. Other kinds of preferences are also accounted for, such as favouring shorter rather than longer words, which typically (but not always) co-incides with the basic-level preference. Reiter notes that humans also employ some preferences that cannot be explained with the parameters investigated so far. He gives the example [Reiter 1991, p. 248] of a speaker pointing the hearer to a cow and a horse with the utterance *Look at the animals / mammals / vertebrates!* None of the terms is basic-level or signigificantly shorter than the others, yet there is a clear order of 'normality' in the sequence of the three candidates.

In my own work on lexical choice in the 'Moose' generator [Stede 1999], I used language-neutral conceptual hierarchies and the subsumption relation, *inter alia* to account for the fact that different languages occasionally display preferences for different levels of specificity. For example, in bi-lingual instructional text we find a regular correspondence between the general English *to remove* and numerous more specific German lexemes (*abziehen, abnehmen, herausdrehen, ...*); this might very well be a genre-specific tendency. Furthermore, Moose employs a model of lexical *connotations* that can override the general preference for a more specific lexical item. For example, when referring to a POODLE in a derogatory manner, Moose can choose the appropriately connotated word *mutt*, which requires moving up the taxonomy to the DOG concept, where a range of near-synonyms (differing in their connotations) are attached. Another reason for considering hyperonyms in the lexical choice process is to avoid repeated usage of the same term when referring to some object multiple times.

In the present Moose implementation, *all* more general words are inherited to the concept-to-be-lexicalized, and the preference mechanism selects one of them (in case of absence of any decisive factors, it chooses the most specific word). This mechanism is certainly not cognitively adequate (it was not intended to be) and also not particularly efficient: The range of candidates under consideration should be constrained beforehand.

In conclusion, NLG systems employ a mixture of *constraints* and *preferences* in their approaches to hyperonymy. The factors used by various systems in the choice process are:

- User's vocabulary and knowledge (e.g., [McKeown et al. 1993])
- Successul reference, i.e., discrimination from other candidate entities (e.g., [Dale, Reiter 1995])
- Basic-level and entry-level effects, conversational implicatures
- Length of words
- Stylistic features such as formality, positive/negative attitude
- Language, genre
- Givenness of item, avoid repetition or "saying the very obvious"

Not surprisingly, there is no generator yet that would incorporate *all* these factors within a single system. It is not clear which general lexical items should be inherited down to the concept-to-be-lexicalized and enter the preferential choice mechanism; it is also not clear how exactly the various preferences would interact and which would take precedence in a particular situation of utterance.

5 Hyperonymy in lexical semantics

Linguists studying lexical semantics are to a good extent concerned with *sense relations* between words, and hyp(er)onymy is certainly one of the relations receiving the most attention. While the intuitive decision whether some entity is subordinate to some other entity is in most cases not difficult to make, spelling out the precise definition of hyponymy (and thus hyperonymy) and its consequences is anything but trivial. Lyons [1977], for example, proposes that *fish* and *bird* share the direct hyperonym *creature* — but not *animal*. That is, when I say *There were plenty of fish in the creek*, the alternative sentence *There were plenty of animals in the creek* would not be a felicitous utterance, even though it is "truth-conditionally correct". And hence, there is a difference between *fish* ISA *creature* and *fish* ISA *animal*.

An interesting distinction in this respect is offered by Cruse [1986], who separates hyponymy from the more constrained relation of *taxonomy*. A diagnosis for the latter is the utterance frame *X is a kind of / type of Y*. Examples that "work" in this frame are: *spaniel–dog*, *rose–flower*, *mango–fruit*. Examples that seem not to work are: *kitten–cat*, *queen–monarch*, *spinster–woman*, *waiter–man*. Notice that both groups are perfectly compatible with the ISA-test, though: No one would doubt that a waiter IS A man, a queen IS A monarch.

Taxonomies, as Cruse proposes, typically have no more than five levels, and frequently have fewer. The levels are commonly labelled as 'unique beginner' – 'life form' – 'generic' – 'specific' – 'varietal'. (The origin of these term in biology is obvious, but they can be transferred to other realms, as Cruse notes.) Most important is the *generic* level, which holds ordinary everyday names like *cat, apple, church, cup*. These items tend to be morphologically simple and are not metaphorically transferred from elsewhere. Most branches of hierarchies terminate at the generic level, and hence this is the level with the largest number of items. Items at specific and varietal levels are particularly likely to be morphologically complex, and compound words are frequent here.

From the notion of explicitly defined levels, it follows that hierarchies do not need to have nodes at each level. Consider the examples in figure 1. Depending on what items people place on the generic level, they end up with one of the two variants; according to Cruse, most people subscribe to the second, which holds *dog, cat, bird* on the same, generic level. Another example are musical instruments: Most of them belong to a kind such as *strings, woodwind, brass, percussion*, but there is no obvious kind for *bagpipes* or *concertina*, which are thus directly linked to *musical instrument*.

Cruse elaborated the importance of the generic level in [Cruse 1977], where he states that for every line of noun taxonomy, there is one term that is 'inherently neutral' (cf. the notion of basic level mentioned above). There is a general rule that requires speakers to use this term in order to obtain an unmarked utterance in a given context — unless this would result in an 'abnormal communication', in which case the speaker should deviate from neutral level, but only to the minimum degree required to ensure normality. Cruse then offers several conditions that would license such over- and underspecification, which we do not reproduce here.

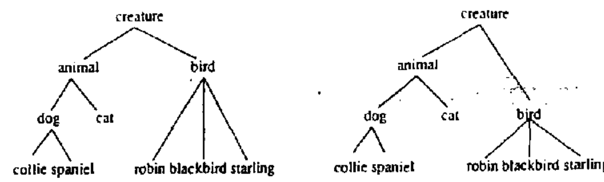

Figure 1: Variants of taxonomy, reproduced from [Cruse 1986, p. 146]

6 Synthesis: Toward a model of conceptual and lexical inheritance

Due to the very different motivations, different kinds of NLG have very different approaches to the hyperonym problem. EI-NLG can basically ignore or finess it. In CI-NLG, it is reduced to a merely technical question: getting the mechanics of spreading activation right, so that lexical convergence enables the subsequent processes of syntactization and articulation (which the CI-NLG models place their emphasis on). A broader view is necessarily based on reasoning with speaker's goals and contextual features, which for the time being is the realm of LI-NLG. Thus, before embarking on building more comprehensive connectionist models, the hyperonym problem is best studied in the frameworks of LI-NLG — but with the motivation of modelling human performance taken into account.

Thus adopting the perspective outlined in section 4, we are interested in choosing words between more or less specific alternatives as well as between near-synonyms of the same specificity. We thereby open the door to both 'vertical' and 'horizontal' lexical choice within a hierarchy, which raises a number of questions:

- What is the granularity of conceptual, and that of lexical knowledge?
- How are the differences between near-synonyms represented?[2]
- Given an activated concept, which more general lexical items are considered in the choice process; are there any restrictions on lexical inheritance?
- How is the eventual choice from the set of candidate lexical items being made?

[2] This question is beyond the scope of this paper; the kind of approach I have in mind here is represented in [DiMarco et al. 1993], [Hirst 1995], [Edmonds 1999].

```
collie -- (a silky-coated sheepdog with a long ruff and long narrow head developed in Scotland)
  => shepherd dog, sheepdog, sheep dog -- (any of various usually long-haird breeds of dog
     reared to herd and guard sheep)
    => working dog -- (any of several breeds of usually large powerful dogs bred to work as
       draft animals and guard and guide dogs)
      => dog, domestic dog, Canis familiaris -- (a member of the genus Canis (probably...
        => canine, canid -- (any of various fissiped mammals with nonretractile claws and
           typically long muzzles)
          => carnivore -- (terrestrial or aquatic flesh-eating mammal; terrestrial carnivores
             have four or five clawed digits on each limb)
            => placental, placental mammal, eutherian, eutherian mammal -- (mammals having a
               placenta; all mammals except monotremes and marsupials)
              => mammal -- (any warm-blooded vertebrate having the skin more or less covered...
                => vertebrate, craniate -- (animals having a bony or cartilaginous skeleton...
                  => chordate -- (any animal of the phylum Chordata having a notochord or
                     spinal column)
                    => animal, animate being, beast, brute, creature, fauna -- (a living
                       organism characterized by voluntary movement)
                      => life form,  organism, being, living thing -- (any living entity)
                        => entity, something -- (anything having existence (living or nonliving))
```

Figure 2: Hyperonyms for *collie* from WordNet

As we have seen, present models that admit hyperonyms into the choice process (in particular those of Reiter [1991] and Stede [1999]) run into the problem of overgeneration: Too many candidates have to be compared for their preferential features, and it is not clear that a decision can always be made.

To illustrate the question of granularity and range of hyperonymic alternatives, contrast the path from *collie* to *creature* given by Cruse [1986] in figure 1 with the hyperonym chain for *collie* offered by WordNet [Fellbaum 1998], shown in figure 2. The WordNet chain includes many items that clearly do not show up in everyday language use, and that a lexical choice process should prefer not to consider when producing an utterance about a collie. *Chordate*, for example, would in the vast majority of utterance situations not be an option. On the other hand, all these terms are certainly 'correct', and a system should be able to respond affirmatively to the question *Is a collie a chordate?*

This divergence points to the need for a distinction between conceptual and lexical granularity and inheritance: The WordNet chain represents rather a series of concepts than of words entering the lexical choice process, which appears to be better represented by a Cruse-type chain with few designated levels (but needs to be augmented with near-synonyms for the 'horizontal' aspects of choice).

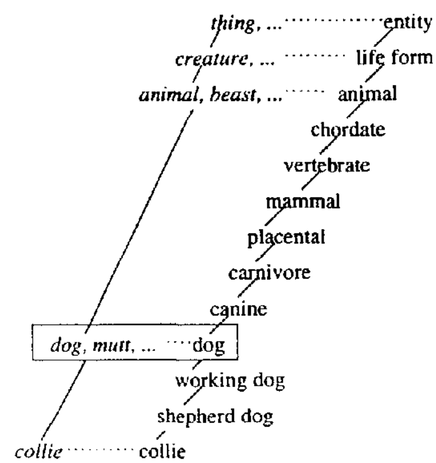

Figure 3: Active-lexical and conceptual hierarchy

The resulting situation is sketched in figure 3. On the right hand side, the nodes of the conceptual chain also are linguistic units, but in language production they would be accessed only if the 'to-communicate' attributes explicitly call for it, e.g., when comparing chordates to vertebrates. Otherwise, only items on the left hand side (tentatively called 'active-lexical') enter the lexical choice process, which are characterized by their particular level in the vocabulary structure, and further differentiated by stylistic and other features. The generic, or basic, level is marked by a box.

When a hyperonym chain is thus not merely an ordered list, but the signficance of the levels is recognized (assuming that Cruse's proposal of level structure indeed scales up to other areas of vocabulary), rules for deviating from the generic level can be stated that map contextual parameters onto 'level movement instructions'. These rules would extend the lexicalisation framework of Reiter [1991], where the first condition is adhering to the hard constraints (the word must convey the essential attributes that are to be communicated), and the second is a preference for the basic level. Adding the instructions for level movement would "contextualize" this framework.

The rules for moving between levels have to consider the specific function of the NP (refer, inform about category membership, etc.) and other factors as indicated in the previous sections (and others mentioned by Cruse [1977]). Since the roles and interactions of these factors are not well understood yet, at this point CI-NLG can make important contributions by designing experiments that shed more light on the parameters that prompt speakers to deviate from the basic level; one example here is the study on speaker's lexical choices in narrative by Downing [1980].

References

R. Brachman, J. Schmolze. "An overview of the KL-ONE knowledge representation system." In: *Cognitive Science* 9 (2), 1985.

D. Cruse. "The pragmatics of lexical specificity." In: *Journal of Linguistics* 13, pp. 153-164, 1977.

D. Cruse. *Lexical semantics*. Cambridge, UK: Cambridge University Press, 1986.

R. Dale, E. Reiter. "Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions." In: *Cognitive Science* 19:233-263, 1995.

C. DiMarco, G. Hirst, M. Stede. "The semantic and stylistic differentiation of synonyms and near-synonyms." In: Working notes of the AAAI Spring Symposium on Building Lexicons for Machine Translation. Stanford University, March 1993.

P. Downing. "Factors influencing lexical choice in narrative." In: W. Chafe (ed.): *The pear stories: cognitive, cultural, and linguistic aspects of narrative production.* Norwood/NJ: Ablex, 1980

P. Edmonds. "Semantic representations of near-synonyms for automatic lexical choice." PhD thesis, Department of Computer Science, University of Toronto, September 1999.

C. Fellbaum. *WordNet — An Electronic Lexical Database.* Cambridge/MA: MIT Press, 1998.

N.M. Goldman. "Conceptual generation." In: R.C. Schank (ed.): *Conceptual information processing.* Amsterdam: North-Holland, 1975.

G. Hirst. "Near-synonymy and the structure of lexical knowledge." In: Working notes of the AAAI Spring Symposium on Representation and Acquisition of Lexical Knowledge. Stanford University, 1995.

J. Lyons. *Semantics. Volume I.* Cambridge/UK: Cambridge University Press, 1977.

K. McKeown, J. Robin, M. Tanenblatt. "Tailoring lexical choice to the user's vocabulary in multimedia explanation generation." In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL).* Columbus, OH, 1993.

W. Levelt. *Speaking: From Intention to Articulation.* Cambridge/MA: MIT Press, 1989.

W. Levelt, A. Roelofs, A. Meyer. "A theory of lexical access in speech production." In: *Behavioral and Brain Sciences* 22, pp. 1-75, 1999.

E. Reiter. "A new model of lexical choice for nouns." In: *Computational Intelligence* 7, 240-251, 1991.

A. Roelofs. "Computational Models of Lemma Retrieval." In: T. Dijkstra, K. de Smedt (eds.): *Computational Psycholinguistics.* London: Taylor & Francis, 1996.

E. Rosch. "Principles of categorization." In: E. Rosch, B. Lloyd (eds.): *Cognition and categorization.* Hilldale, NJ: Lawrence Erlbaum, 1978.

M. Stede. *Lexical semantics and knowledge representation in multilingual text generation.* Dordrecht/Boston: Kluwer, 1999.