WordNet Nouns: Classes and Instances

George A. Miller
Princeton University

Florentina Hristea
University of Bucharest

WordNet, a lexical database for English that is extensively used by computational linguists, has not previously distinguished hyponyms that are classes from hyponyms that are instances. This note describes an attempt to draw that distinction and proposes a simple way to incorporate the results into future versions of WordNet.

If you were to say "Women are numerous," you would not wish to imply that any particular woman is numerous. Instead, you would probably mean something like "The class of women contains numerous instances." To say, on the other hand, "Rosa Parks is numerous," would be nonsense. Whereas the noun *woman* denotes a class, the proper noun *Rosa Parks* is an instance of that class. As Quirk et al. (1985, page 288) point out, proper nouns normally lack number contrast.

This important distinction between classes and instances underlies the present discussion of WordNet nouns. Some nouns are understood to refer to classes; membership in those classes determines the semantic relation of hyponymy that is basic for the organization of nouns in WordNet (WN). Other nouns, however, are understood to refer to particular individuals. In many cases the distinction is clear, but not always.

The distinction to be discussed here is between words ordinarily understood as referring to classes and words ordinarily understood as referring to particular individuals and places. In the literature on knowledge representation, the classic discussion of this distinction is provided by Woods (1975). The distinction was not drawn in initial versions of WN (Miller 1990; Fellbaum 1998), which used the "is a" relation in both cases. That is to say, both "A heroine is a woman" and "Rosa Parks is a woman" were considered to be occurrences of the "is a" relation and were encoded in the WN database in the same manner.

Requests to incorporate a distinction between classes and instances have come from ontologists, among others. In their discussion of WN, for example, Gangemi et al. (2001) and Oltramari et al. (2002) complain about the confusion between individuals and concepts. They suggest that if there was an "instance of" relation, they could distinguish between a concept-to-concept relation of subsumption and an individual-to-concept relation of instantiation. That is, essentially, the suggestion we follow in the present work, but in some cases the distinction was not easy to draw.

Incorporating this distinction was resisted at first because WN was not initially conceived as an ontology, but rather as a description of lexical knowledge. WN includes verbs, adjectives, and adverbs in addition to nouns. Although no ontology was intended, the organization of nouns in WN bore many similarities to an ontology. As the importance of ontology became more apparent, requests to convert the WN noun hierarchy could no longer be ignored. Version 2.1 of WN takes a step in that direction: the Tops file is reorganized to have a single unique beginner: *entity*. In a reasonable ontology, however, all terms might be expected to conform to the membership relation of set theory and would not contain individuals or placenames. The confounding of classes and instances in WN posed a problem. The obvious way to solve

that problem was to distinguish them. That is to say, the instances in WN had to be identified.

There are three characteristics that all words denoting instances share. They are, first of all, nouns. Second, they are proper nouns, which means that they should be capitalized. And finally, the referent should be a unique entity, which implies that they should not have hyponyms; it is meaningless to have an instance of an instance.

Unfortunately, these characteristics are shared by many words that are not instances. In clear-cut cases, such as persons or cities, there is little problem identifying instances. But there are many other proper nouns that are not instances. It was decided that there was no alternative to inspecting the 24,073 sets of synonyms that contained candidate nouns, one at a time. This was done manually by the authors, FH and GM.

The strategy agreed on for assigning "instance" tags was to concentrate on a word's referent. When they knew of a unique referent, it was considered a clear case of an instance. Otherwise it was considered a class. For example, when *Beethoven* is used to refer to the German composer, it is an instance, but when *Beethoven* is used to refer to the composer's music (as in "She loved to listen to Beethoven") the same word refers to a class of musical compositions. Moreover, just to be clear, when there were two different referents, both were tagged as instances. For example, *Bethlehem* in the Holy Land and *Bethlehem* in Pennsylvania were both tagged as instances. And when an instance had two or more hypernyms, it was tagged as an instance of all of them. For example, *Mars* is an instance of a terrestrial planet (having a compact rocky surface) and also as an instance of a superior planet (its orbit lies outside the Earth's orbit).

The basic entries in WN are sets of synonyms, or synsets. A problem reported by both taggers was the occurrence of capitalized and lower-case words in the same synset. It makes no sense for a word to refer to an instance and for its synonym to refer to a class, so in these cases the entire synset was considered to denote a class. For example, *acetaminophen* and *Tylenol* are synonyms in WN and both were considered to denote classes. The possibility that *Tylenol* might be an instance of *acetaminophen* seemed to be refuted by such usages as "She took two Tylenol and went to bed." In short, giving something a trade name does not change it from a class to an instance. The street names of drugs were also considered to denote classes.

The two taggers disagreed in their treatment of sacred texts. Whereas they agreed that *Adi Granth, Zend Vesta, Bhagavadgita, Mahabharata,* and others were instances of sacred texts, when they came to the Christian *Bible* they disagreed. FH considered it an instance, no different from other sacred texts; GM called it a class term because there are many hyponyms of *Bible: Vulgate, Douay, King James, Revised Version, American Revised Version,* etc. But GM's decision made the Bible a special case, which may have resulted from WN's compilers knowing more about the Bible than they knew about other sacred texts. It was decided that this was a case in which a sacred text could be a class: *Bible* was tagged as a class of a sacred text and its hyponyms were tagged as instances.

Languages posed another problem. For example, are *Old Italian, Sardinian,* and *Tuscan* instances of *Italian?* It was decided that, from an ontological point of view, languages are not instances. Only speech acts are instances.

Placenames included many geographical regions that do not have well-defined political boundaries: *Andalusia, Appalachia, Antarctic Zone, Badlands, Barbary Coast, Bithynia, Caucasia,* etc., but the terms still have geographical significance and are in general use. Although vague in denotation, the taggers considered them instances. The names of cities and states, islands and continents, rivers and lakes, mountain peaks and mountain

ranges, seas and oceans, planets and satellites, stars and constellations, were tagged as instances, as were the signs of the zodiac.

The names of numbers and the names of monetary units were all considered as classes; the *Hong Kong dollar,* for example, is not an instance of *dollar.*

Overall, there were 7,671 synsets in WN that the taggers finally agreed should be tagged as instances. Version 2.1 of WordNet contains these distinctions and it will be subjected to helpful criticism by WN users, as are all the other lexical relations in WN.

Finally, a word about the notation that will represent this distinction in WN. The symbol used to code hypernyms has been @. That is to say, {*peach, drupe,*@} has represented "a peach is a drupe" or "all peaches are drupes." This notation is appropriate for representing relations between classes but it is not appropriate for representing relations of instantiation. When {*Berlin, city,*@} is used to represent "Berlin is a city," the instance *Berlin* is treated inappropriately as a class. A different symbol is needed to code instances. It has been decided, therefore, simply to add an i to the @; to represent "Berlin is an instance of a city" by {*Berlin, city,*@i} in the new notation.

Since WN 2.1 contains the distinctions between classes and instances described here, it will be possible to treat WN nouns as a semi-ontology by simply ignoring all nouns tagged with @i. It is hoped that this modification will make WN more useful to future users.

Acknowledgments

Florentina Hristea is grateful to the Romanian-U.S. Fulbright Commission for the Fulbright Grant that made it possible for her to collaborate in this research. Work by the Cognitive Science Laboratory was supported by a contract between Princeton University and the Advanced Research Development Activity (AQUAINT Program Phase 2, Contract No. NBCHC40012). The authors are indebted to Benjamin Haskell for developing the interface that was used to tag instances and to Christiane Fellbaum, Helen Langone, and Randee Tengi for comments on the manuscript.

References

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Gangemi, Aldo, Nicola Guarino, and Alessandro Oltramari. 2001. Conceptual analysis of lexical taxonomies: The case of WordNet top-level. In C. Welty and B. Smith, editors, *Formal Ontology in Information Systems*. *Proceedings of FOIS2001.* ACM Press, 285–296.

Oltramari, Alessandro, Aldo Gangemi, Nicola Guarino, and Claudio Masolo. 2002. Restructuring WordNet's top-level: The OntoClean approach. In *Proceedings of LREC2002* (OntoLex workshop), Las Palmas, Spain.

Miller, George A., editor. 1990. WordNet: An on-line lexical database [Special Issue]. *International Journal of Lexicography*, 3:235–312.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, London and New York.

Woods, William A. 1975. What's in a link: Foundations for semantic networks. In Daniel G. Bobrow and Alan Collins, editors, *Representation and Understanding: Studies in Cognitive Science.* Academic Press, New York.