Building a Large Lexicon with Lexical Network Theory

Peter Norvig*
University of California at Berkeley
Computer Science Division, Evans Hall
Berkeley, California 94720 USA
norvig@cogsci.berkeley.edu

Abstract

In this paper I present a number of studies of individual lexical items demonstrating the incredible variety of idiosyncratic properties that occur in the lexicon of English. While many current theories ignore this level of detail, I argue that a complete on-line lexicon suitable for a variety of linguistic and computational uses must include much more information than is available in any existing dictionary. I present a formalism, Lexical Network Theory (LNT) which helps deal with the complexity by capturing generalizations both about individual word senses and about the relations or links between word senses. In addition, a methodology for building lexical networks is proposed. This methodology manages the interaction between the lexicographer, a large corpus of word uses, and the evolving lexical net so that the complexity of the net can be used as an asset to suggest interpretations for each new word, rather than as a liability that overwhelms the lexicographer by its sheer size.

1 Introduction

Lexical Network Theory (LNT) asserts that the semantic portion of the lexicon is best seen as a network of word senses, where each sense is connected by links to other semantically-related senses of the same word, and, indirectly, to other words in the same semantic field. In other words, a link is a function which operates on one sense to produce another. Characterizing the links is as important as characterizing the individual word senses. In fact, the links may be more important, in that they provide the generalizations that allow us to build up a large lexicon. Using the function analogy, we can build a lexicon faster by defining some representative senses and general functions operating on those senses to produce a large number of new senses, as opposed

*Supported by the Defense Advanced Research Projects Agency (DoD), Arpa Order No. 4871, monitored by Space and Naval Warfare Systems Command under Contract N00039-88-C-0292.

to defining each sense explicitly. The problem is that lexical generalizations are not as clear-cut as mathematical functions.

The LNT approach is in contrast to theories where the lexicon is a flat list of syntax/semantics pairs (or phonology/subcategorization/semantics triplets), and where lexical redundancy rules serve only as a convenient notation for a larger list of lexical items.

One way of viewing Lexical Network Theory is that we are simply elevating the notion of a lexical redundancy rule from an extra-theoretic notation device to a formal part of the theory. In a broad sense, this is a correct characterization. However, the links are also used to express metaphorical, metonymic, and polysemy relations. Sequences of links can express synonymy, antonymy, hyponymy and other inter-word relationships that are not expressible by lexical redundancy rules. This is because LNT insists that the semantic frame(s) associated with a sense be fully specified, and related to other senses and other words, while the methodology associated with flat-list lexicons with redundancy rules normally uses unanalyzed predicates for semantics, rather than fully-analyzed semantic frames.

2 Advantages of LNT

There are four places where LNT can play a useful role in building large lexicons:

Intra-word. LNT provides a convenient way to characterize multiple senses of a word, and a means of capturing generalizations by stating new senses in terms of differences from existing senses, rather than by stating a completely new, separate definition.

Inter-word. Knowing that word a has sense a_1 and related sense a_2, and that word b has sense b_1 similar to a_1, we can predict that b may also have sense b_2, which is derived from b_1 in the same way that a_2 is derived from a_1.

Inter-language. Knowing that word a in language L_1 corresponds to word α in language L_2, we can predict senses of α based on the senses of a, and knowledge of which link types from L_1 are valid in L_2.

Historical. If word sense a_3 is related to sense a_1 by some complex link l_3, that is evidence that there may have been a sense a_2 which is related to a_1 by l_1 and to a_3 by l_2, but which is no longer present in the language. The composition of l_1 and l_2 yields l_3, but we prefer analyses with simple links only.

In addition, because the links are formal objects in the system, they can be made available to other processes outside the lexicon proper. For example, a parser could compute meanings for novel word uses on the fly. “The paperboy porched the paper” (see [Clark and Clark, 1979]) could be handled by combining a known denominal-forming link with contextual knowledge about newspaper delivery and porches. The ability to do on-the-fly lexical analysis is incompatible with approaches that see the lexicon as a flat list.

3 Methodology

Fillmore and Kay [1989] have proposed a methodology for developing lexicons based on their Construction Grammar formalism, and on some of the tenets of LNT. Here I outline a slightly more ambitious version of their methodology, supplemented by the support of automated tools based on LNT. At the outset, let it be clear that the project is a difficult and ambitious one, due to the need for detailed semantic analysis, and the commitment to cover the complete language as it is used, rather than being limited to "core" constructions and lexical senses only. The current plan is to work with an available corpus of approximately 100 million words.

The first step in the methodology is to select a particular semantic domain to be examined, along with the associated terminology. For example, the domain of kinship terms includes the terms "brother," "mother-in-law," "second cousin," and "step-grand-daughter." The domain of commercial transactions includes the terms "buy," "sell," "pay," "spend," "charge," "cost," "pawn," "merchandise," and "price."

The selected domain must then be analyzed and described. We strive for a compact, formal description in terms of logical relations whenever possible. For example, in the kinship domain, this is readily achievable. In other domains formal description may be augmented or replaced in part by other *ad hoc* description devices. For example, in the domain of body part terms, a labeled diagram would prove very useful. The descriptions serve three audiences: the automated acquisition routines, the analyst, and the eventual human user of the lexicon. We want to include all uncovered facts, even if the analysts have not yet decided on the proper representation for some information, so we allow for informal annotation as well as formal descriptions.

Next, we examine the usages of each lexical item in the semantic domain. For each word, we extract a concordance from the corpus, and then label each usage. Initially, this has been a mostly manual task, with limited automation. For example, usages of the word "risked" can be automatically classified as verbs, but "risk" and "risks" must be manually classified as either verb instances or noun instances. As coverage increases, some classification can be done by a parser, rather than by relying on a morphological analyzer alone. Each word usage must also be subcategorized for the complements it takes (and for the complement position in which it appears). In addition to syntactic properties, this subcategorization must take into account the semantic role played by each constituent according to the domain semantics defined above. This entails a much more fine-grained analysis than any existing theory of thematic roles can provide.

The results of this analysis must be organized into a network of senses for each word, where each analyzed usage is classified under some sense. Fillmore's analysis of "risk," for example, lists three senses of the verb with nominal object: the direct object can be the valued object that may be harmed or lost (He risked *his life*), it can be the harm itself (He risked *death*), or it can be the action which leads to the threat (He risked *a trip into the jungle*). It may also be a metonymy for that action (He risked *the jungle*). (It is not clear if this should be a fourth sense, or if it is a general metonymy that is not a purely lexical property of "risk.") There are also two main usage classes

with gerundial complements: harm (She risked *losing her job*) and action (She risked *hiring an unproven worker*).

Of course, during the course of the analysis it will be typical to see changes to the classification and subclassification, to the analysis of the semantic domain, and even to to the list of relevant terms, as the analyst(s) gains more experience and insight into the domain. Fillmore and Atkins' study of "risk" involved over 1700 usages, and they have not yet undertaken a thorough study of related terms such as "jeopardy/jeopardize" and "hazard." The major difficulty is just keeping track of this vast array of data, and it is here that Lexical Network Theory can help. Given a partially specified lexical network, LNT can help the lexicographer analyst in the following ways:

- Based on the subcategorization frame of a word, suggest that the usage may be another instance of a previously established sense of this word. Previously classified usages of that sense can be displayed on request.
- Suggest that the usage may be a new sense, derived by application of a link that had been defined for some other word sense.
- Given a proposed classification under some previously established sense, inform the lexicographer that there is an inconsistency between the usage and the sense, or confirm that the proposed classification is consistent. In the case of an error, the system can show the problem, and suggest alternative classifications.
- Given a proposed new sense, the system can inform the lexicographer that the analysis is the same as a previously established sense, and ask if that is the intended sense, or if there is some additional information that distinguishes the two.
- For low-frequency lexical items, we can have the system apply known links to generate plausible sentences, and then ask the analyst for an acceptability judgment. This technique can also be used for high-frequency words to test the boundaries of some link: we can apply a known link with uncertain applicability conditions to new words, thereby sharpening the applicability conditions. We note that negative judgments in response to generated examples are more helpful than positive responses, since negative judgments help define the boundaries of a sense, but there are by definition no negative instances in the corpus.

None of this would be possible in an informal system that did not define explicit links between senses.

Although we are still in the prototype stage, we recognize a potential problem: for large lexicons, the number of suggestions may become prohibitively large. We believe this problem can be alleviated by training the system, through the use of feedback by the lexicographer. In addition to accepting or rejecting the system's suggestions, we will offer the option of commenting on the appropriateness of the suggestion. When a suggestion made by a particular link has been consistently marked inappropriate, the system will refrain from making suggestions based on that link.

Note that acceptable senses can still be marked as inappropriate. For example, suppose the lexicographer has explained the usage "The girl in the picture has green eyes" by adding a new sense for "girl" whose semantics is "an iconic representation of a girl." This is connected by a metonymic link to the primary sense of "girl." When the lexicographer subsequently adds a new noun, the system may suggest the iconic representation sense. The lexicographer can judge these suggestions as acceptable but inappropriate, and eventually they will stop being suggested. The details of this mechanism have not been worked out.

4 Examples

In this section we cite a few examples of lexical studies done within the Construction Grammar/LNT framework, in an effort to display the variety of information that must be accounted for. Note that none of these studies made use of the kind of automated tools we are proposing in this paper, yet each study resulted in the discovery of some novel lexical facts that had not been previously noted in dictionaries or in published papers.

4.1 Take

As our first example, Norvig and Lakoff [1987] examined the verb "take," including usages like the following:

1. John took the book from Mary.
2. John took the book to Mary.
3. John took the book to Chicago.
4. John took Mary to the theater.
5. John took Mary to class.
6. John took a whiff of the coffee.
7. John took a punch from Harry.
8. John took a punch at Harry.

There are some puzzling questions involved in describing these data:

- Are these eight different senses, or less than eight?
- Why is it that the subject can be stationary in 1, but not in 2 or 3?
- Why must there be change of possession in 2, but not 3?
- Why does 4 imply both John and Mary attended, while 5 does not?
- Why can "take" express perception, as in 6?
- Why is the direction of action reversed in 7 and 8?

Abstractionist definitions, which define "take" as "cause to have" or something similar, fail to account for the variety of the data. Norvig and Lakoff [1987] are able to define a network of seven senses that account for all the data above, as well as other uses. (There are only seven senses because 3 and 5 end up the same sense.) In addition, we can explain much about the related verbs "gives," "gets," "bring," "make" and "has," by recourse to the link-types defined for "take." However, we are still a long ways from a complete description of any of these words.

Each sense is related to another by a simple link describing a single difference. It is a principle that links can be composed only of simple, single differences, along with any differences that are thereby entailed. It is conjectured that more complex links do not occur because they would be too difficult to learn. Of course, we do not rule out the possibility of homonyms, where senses of the same word are not linked at all to each other.

4.2 Over

A systematic use of LNT can help the lexicographer make difficult decisions about the difference or identity of word usages. The study of "over" described in [Brugman and Lakoff, 1988] considered many usages, including ones like the following:

fly over the yard (+X,-V,-C)
fly over the hill (+X,+V,-C)
fly over the spot (-X,-V,-C)
fly over the wall (-X,+V,-C)
walk all over the yard (+X,-V,+C)
walk over the hill (+X,+V,+C)
sit over the spot (-X,-V,+C)
climb over the wall (-X,+V,+C)

Here the $\pm$X refers to objects that are horizontally extended or non-extended, $\pm$V refers to a vertical or non-vertical object, and $\pm$C refers to contact or non-contact between subject and object. From the examples given, there is no reason to assume that these are salient properties, and thus one could reasonably assume that all eight uses refer to the same sense. However, when we consider senses where the focus is on the endpoint of the path (marked +E), we see that some of the properties come into play:

He lives over the bridge (+X,-V,+E)
He lives over the hill (+X,+V,+E)
* He lives over the spot (-X,-V,+E)
* He lives over the wall (-X,+V,+E)

With endpoint focus, "he lives over the bridge" means that he lives at the end of the path that traverses the bridge, and is thus *not* complementary to "He lives under the bridge." In contrast, "He lives over the spot" does not accept the endpoint focus reading, and can only mean that he lives in the space above the spot. In summary, only +X senses accept the +E focus. This legitimizes the use of X as a feature (and other facts legitimize the other features), but it does not force an analysis of the first eight senses listed above as wholly distinct. Brugman prefers an analysis where those uses are simultaneously

sanctioned by one of the eight senses, and by a general abstract sense meaning roughly "above and across."

Note that many of the links presented in this analysis are useful with other lexical items as well. The endpoint focus link, for example, is widely applicable. Consider these (-E) examples:

Sam walked *over* the hill.
Sam walked *around* the corner.
Sam walked *across* the street.
Sam walked *down* the road.
Sam walked *past* the post office.

as compared with the (+E) examples:

Sam lives *over* the hill.
Sam lives *around* the corner.
Sam lives *across* the street.
Sam lives *down* the road.
Sam lives *past* the post office.

On the basis of these examples, we can propose a general endpoint focus link, which applies to any sense dealing with motion along a path, and returns a new sense that keeps all the same relationships, but refers to the endpoint of the path, rather than the path itself. However, this applicability condition is not quite strict enough. For example, "by" has no endpoint meaning: "Sam lives *by* the post office" means that he lives near there, not that he lives at the end of the path that goes by the post office. Thus, while it is important to seek general principles, it is equally important to notice exceptions. So, according to our methodology, when the word "by" is considered, the system would suggest the application of the endpoint focus link, and the analyst would reject the suggestion (but nevertheless mark it as an appropriate suggestion to have made). After several such rejections, the analyst may have a better idea of the applicability conditions of the link, and could make an explicit change.

This example also points out that the methodology in practice is not quite as neat as portrayed in Section 3 above. The methodology suggests that the analyst defines the semantic domain before cataloging the data. However, there will always be cases where a usage suggests a new extension to the domain: it would have taken great foresight indeed to include a notion of endpoint focus in an initial analysis of the semantic domain for "over." Thus, the methodology given in Section 3 defines just one round in what could be a large number of iterations needed to achieve the final analysis.

4.3 Have and With

Brugman's [1988] in-depth study of "have" reveals it to be the most syntactically promiscuous verb in the language (even more so than "be"), subcategorizing for at least eleven different syntactic frames. In the listing below, they are named by the number of complements (excluding the subject), and the form of the final complement:

2VP$_{en}$: I had my bicycle stolen.
2VP$_{ing}$: I had him climbing the walls.

2VP$_{inf}$: I had him bring chips to the party.
2AP: I have a tooth missing.
2PP: I have $5 in my pocket.
2NP: The Times had Bush (as) the winner.
2VP$_{to}$: I have the cat to feed.
2S$_{that}$: Rumor has it (that) he quit.
1VP$_{en}$: We have eaten already.
1VP$_{to}$: I have to go to work.
1NP: I have $5.

In addition to this syntactic classification, there are five major semantic types:

Attributive: I have two siblings.
Affecting: I had my bicycle stolen.
Resultant: I had him climbing the walls.
Causative: I had him bring chips to the party.
Depictive: The Times had Bush (as) the winner.

There is no direct mapping between the five semantic and eleven syntactic types, although not all of the 55 combinations occur. The full analysis of "have" describes which combinations are possible, through a combination of general principles and explicit lists. Rather than go into detail here, I will turn to the connection with the preposition "with," which shares some meaning with "have." It turns out that "with" also shares many of the subcategorization frames:

2VP$_{en}$: With my bicycle stolen, I walk.
2VP$_{ing}$: With him quitting, we're shorthanded.
2AP: With the boss sick, we got the work done.
2PP: With $5 in my pocket, I can pay for it.
2NP: With Bush the winner, we will suffer.
1NP: With $5, I can pay for it.
2VP$_{to}$: With the cat to feed, I'll be busy.

However, some of the constructions fail to work with "with":

2VP$_{inf}$: * With him bring chips, we were covered.
2S$_{that}$: * With it (that) he quit, we're shorthanded.
1VP$_{en}$: * With eaten already, I'm not hungry.
1VP$_{to}$: * With to go to work, I can't see you.

On the semantic side, we can make the generalization that "with" cannot be used to represent the causative case, but works in all the other cases. This study shows that there is substantial overlap between the uses of "have" and "with," and that the two words might better be studied together, rather than in isolation. By relating the two to each other, we can both make the analysis process easier, and arrive at cleaner, more general final analysis.

4.4 Meat

We need a better theory of what can constitute a valid link between senses. As an example of the problem, consider the class of links that expand a central, prototyp-

ical sense to include peripheral interpretations. These have been studied by Stallard [1987], among others. He points out that the noun "ship" is polysemous between a central sense that excludes submarines, and an extended sense (in this case, the more "technical" sense) that includes both submarines and surface ships. (There is a third sense that includes air- and space-ships as well.) Thus, both of the following are felicitous statements:

That's not a ship_1, it's a submarine.

Of course that's a ship_2; it's a submarine.

This type of polysemy is very common. We could replace the ship/submarine pair in the paradigm above with lion/lioness, animal/human, blue/turquoise, unwise/insane, or dark/black, for example.

Unfortunately, just noting these semantic extensions does not tell the whole story. It turns out that the exact extensions, and the words they operate on, are very idiosyncratic. Consider the case of the word "meat." The central sense is something like: (1) "the edible muscle tissue of a mammal (especially a bovine), when stripped from the animal and intended for consumption." There are a number of directions in which extensions are admitted:

(2) Allow fowl as well as mammals.

(3) Allow fish as well as fowl and mammals.

(4) Allow organ meat as well as muscle tissue.

(5) Allow skin as well as muscle tissue.

(6) The edible part of any food (e.g. coconut meat).

(7) (Metaphor) The core or essence of something.

Sense (3) is derived from sense (2), while the rest are all derived directly from sense (1). That these polysemous senses exist can be seen by examples like the following:

Bring me chicken, not beef; I don't eat meat.

Bring me fish, not chicken; I don't eat meat.

Bring me rice, not fish; I don't eat meat.

* Bring me chicken, not fish; I don't eat meat.

* Bring me pork, not beef; I don't eat meat.

The last two examples demonstrate the fixed ordering of senses 1,2,3. Suffice it to say that while these examples can all be classified as instances of extension from a central prototype, it is clear that no *a priori* model of extension, and only a very precise model of *meat* as it is conceived in our culture could account for the exact variety of extensions.

4.5 Play

In some cases it is difficult to know where the lexicon ends and the grammar begins. In the combined Construction Grammar/Lexical Network Theory this is not a vital concern, since lexical entries are just special cases of constructions; there is no hard distinction. This makes it possible to define general metaphors and metonymies that are explicitly applied to individual words, but can also be applied on the fly to novel expressions, or to anaphors with the proper referent.

One place where the lexicon and grammar come into contact is the assignment of peripheral prepositional phrases. Most theories hold that in a sentence like "John went to NY on Tuesday," the verb "went" subcategorizes the PP_{to}, but the PP_{on} is attached at the sentence level, and is not subcategorized by "went." A study of the verb "play" indicates that there can be idiosyncratic semantic interactions between the verb and such sentence-level PPs that are reflected in the syntactic ordering of the PPs. We will concentrate on this below, to the exclusion of many other interesting facts about "play."

In the *make music* sense of "play," there are two main scenarios. In the first, a musician manipulates a musical instrument to produce sounds representing a piece conceived by some composer, so that the sounds are heard by the audience. In the second scenario, the sounds are recorded on some physical media, such that they can be reproduced later over some broadcast media that typically includes some electronic transmission followed by the vibration of a speaker. The net result is again that an audience hears the sounds. It is an interesting fact about "play" that it can refer to either of these scenarios, or both at once:

Stern played at Carnegie Hall today.

The radio played the emergency test warning signal.

The radio played a recording of Bach's concertos.

"Play" admits a large number of subcategorization frames. For example, the subject can be the musician, the piece, the composer, the musical instrument, the recording, the transmission media (e.g. radio) or the reproduction media (e.g. speakers). All of these can also appear in object position, with a suitable subject. However, for the current point, the interesting case is when these roles are represented by PP_{on}. In these cases the ordering of PPs is not free, but prefers an iconic ordering corresponding to the flow of force/sound from musician through instrument to recording media and eventually to the audience. That is, the flow of force in scenario one is from composer and musician (taken together) to musical instrument to audience. In the second scenario, it is from song to recording media to transmission media to reproduction media (e.g. speakers) to audience. The most felicitous, unmarked cases follow these paths in the left-to-right ordering of complements:

A child is playing on the piano on this tape.

The radio is playing on these speakers.

He played flute on the synthesizer on that record.

A piano is playing on the radio on these speakers.

The examples below violate the ordering and are less acceptable or unacceptable:

*This tape is playing on the piano.

*These speakers are playing on the radio.

?He played flute on that record on the synthesizer.

*A piano is playing on these speakers on the radio.

5 Complications

So far we have described the lexical network as a graph which is a forest: each distinct entry is the root of a tree, and semantically related senses branch off of the root. There are at least four complications that need to be explained.

First, sub-networks for distinct words can be indirectly linked, by way of an intervening semantic representation. For example, we would link both "come" and "go" to the movement semantic frame, and thus they are indirectly linked to each other. Of course, they are not synonymous, because they have different restrictions on the use of the frame.

Second, some senses have multiple motivations, yielding a directed graph structure instead of a tree. The lexicographer can volunteer two mutually supporting links to a single new sense, or it may be that he or she proposes a single derivation, but the system proposes another that the lexicographer can then either accept as being equally accurate or reject. We have found that it is necessary to store negative links (x is not derived from y by link L) to avoid having to reject the same proposed derivation over and over again. An example of multiple motivation, presented in [Martin, 1988] is the sense of "communicate" as in "the flu can be communicated by sneezing." This seems to be derived both from the infection-as-transfer metaphor ("I gave him my cold") and the communication-as-transfer metaphor ("I got my ideas across to her.").

Third, there can be gaps in the network due to historical loss. What appears to be two distinct trees associated with two homonyms may have at one time been a single tree linked by an intervening sense. One possible strategy is to take all generated candidate senses with multiple motivations and show them to a historical linguist. They might represent historically lost senses.

Although LNT is much more than lexical redundancy rules, we are subject to some of the same difficulties that plague redundancy rules. For example, we need a satisfactory account of ordering constraints to differentiate between the analysis where sense a_3 is derived from a_2 by l_2 (where a_2 is derived from a_1 by l_1) as opposed to the analysis where a_3 is derived from $a_{2'}$ by l_1 (where $a_{2'}$ is derived from a_1 by l_2). We do have the advantage that, in cases where both a_2 and $a_{2'}$ are valid senses, we can simply say that both derivations are equally valid. This is no contradiction because lexical networks represent a static description rather than a dynamic derivational process.

Just as some senses can be predicted from general principles, it is also possible for some senses to be ruled out by general principles acting together with the rest of the lexicon. [Clark and Clark, 1979] give the example of *preemption* as it applies to some fairly general (but not completely productive) denominalization links. For example, one can say: "He *bicycled/skateboarded/helicoptered* ten miles" to describe motion using a particular vehicle. However, conspicuously missing from this paradigm are the uses "He **carred/*airplaned* ten miles," presumably because they are preempted by the common verbs *drove* and *flew*. The absence of the verbs "carred" and "airplaned" should not be listed as a fact about the nouns "car" and "airplane," but rather as a fact about the verbs "drove" and "flew," together with the preemption principle.

Acknowledgments

I am indebted to Claudia Brugman, Charles Fillmore, Paul Kay, and George Lakoff for useful discussions, examples and advice, and to Dekai Wu for proofreading and stylistic changes.

References

[Brugman, 1988] Claudia Brugman. The syntax and semantics of HAVE and its complements. Ph.D. thesis, Linguistics, University of California, Berkeley, 1988.

[Brugman and Lakoff, 1988] Claudia Brugman and George Lakoff. Cognitive topology and lexical networks in Small, Cottrell, and Tanenhaus, eds., *Lexical Ambiguity Resolution*, Morgan Kaufman, 477-508, 1988.

[Clark and Clark, 1979] Eve V. Clark and Herbert H. Clark. When nouns surface as verbs. *Language*, 55(4):767-809, 1979.

[Fillmore and Kay, 1989] Charles Fillmore and Paul Kay. Computational Lexicography in Construction Grammar. ms., 1989.

[Martin, 1988] James Martin. A computational theory of metaphor. Report No. UCB/CSD 88/465, Computer Science Division, University of California, Berkeley, 1988.

[Norvig and Lakoff, 1987] Peter Norvig and George Lakoff. Taking: a study in lexical network theory. In *Proceedings of the 13th Berkeley Linguistics Society Annual Meeting*, pages 195-206, Feb. 1987.

[Stallard, 1987] David Stallard. The logical analysis of lexical ambiguity. In *Proceedings of the 25th Annual Meeting of the ACL*, pages 179-185, July 1987.