Using an On-line Dictionary to Extract a List of Sense-Disambiguated Synonyms

G. Jan Wilms
Computer Science, Mississippi State University
WILMS@CS.MSSTATE.edu JGW1@MSSTATE.bitnet
POBox 1715, Mississippi State, MS 39762

ABSTRACT

The feasibility of extracting both explicit and implicit synonym references from a machine readable dictionary is investigated; the extracted synonyms, both symmetric and asymmetric, are then sense-disambiguated. At the same time lemma numbers and unbound parts-of-speech of synonyms become instantiated. The dictionary source is also a resource for parsing the definitions, but its comprehensiveness is often a mixed blessing as a disambiguation tool.

INTRODUCTION

Most computational lexicons used in natural language understanding systems have been manually constructed, a painstaking effort that is error-prone, labor intensive, and usually is attempted only to create a bare-bone lexicon that is sufficient for the immediate operation of the program when applied to a restricted domain of interest. To overcome this bottleneck many researchers have turned to machine readable resources as a possible source for automating the acquisition of a semantic lexicon; various taxonomies of semantic relations have been extracted (e.g. AMSLER 81, CALZOLARI 84, BYRD et al. 87, VERONIS and IDE 90).

This paper investigates the feasibility of automating the extraction of a list of sense-disambiguated synonyms from a machine readable dictionary (*the Funk and Wagnalls (F&W) Dictionary*). Since most words are polysemous, identifying what sense a synonym is used in is essential to avoid relating words based on orthographic in stead of semantic similarities (e.g. *disorder* is a synonym of *sickness*, but only in the sense of *ailment*, not in the other senses of *confusion*, or *tumult* and *riot)*. Chodorow used the synonyms and hypernyms in the on-line version of *The New Collins Thesaurus* to compute the semantic distance between any two words (using a process he calls *sprouting*), but found that even after sense-disambiguating the thesaurus many words remained spuriously related because of "poor sense separation" in the thesaurus (CHODOROW 88, p. 149). This paper argues that a dictionary has better sense separation than a thesaurus.

This work was supported in part by the NSF grant number IRI-9002135. I would like to thank Lois Boggess for her helpful comments.

The extracted taxonomy of related words will be used in an ongoing project at Mississippi State University, that is attempting to automate, in a domain-independent way, the extraction of knowledge contained in a machine readable technical corpus into an object-oriented knowledge base (HODGES et al. 91). An important research issue involves checking for redundancy; i.e. when more than one name is used to refer to a single real-world object (e.g. *hemorrhage* and *bleeding*), they should be mapped to only one knowledgebase object. Another issue is the automatic bootstrapping of a semantic lexicon for the domain text; one strategy for dealing with unknown words is to check for semantically related entries that already exist in the lexicon.

FUNK AND WAGNALLS STANDARD DESK DICTIONARY

In the preface to the *F&W* dictionary, synonyms are defined in terms of verbal equivalences: "what word or phrase or circumlocution can serve as an equivalent for the word defined" (p xxii). The test for synonymy is given as interchangeability (in certain contexts), and the entry for synonym in the dictionary defines it as "a word having the same or almost the same meaning as some other: opposed to antonym". Synonyms are explicitly marked with the keyword **Syn.** Sometimes there is a pointer to a particular sense of the lemma being defined (e.g. *Alien[1]: adj. 1. Owing allegiance to another country; unnaturalized; foreign. 2. Of or related to aliens. 3. Not one's own; strange. 4. Not consistent with; incongruous; opposed: with to. — n. 1. An unnaturalized foreign resident. 2. A member of a foreign nation, tribe, people, etc. 3. One estranged or excluded. — **Syn. (adj.) 4. extrinsic, extraneous, irrelevant.*** The superscript [1] next to the headword (the lemma being defined) is a lemma number; see below).

	Number of synonyms recovered	% of total
Explicit References	898	2.3
'Also'	925	2.4
Collateral Adjectives	95	0.3
One-word definitions	30,720	79.7
Conjoined one-word definitions	5,526	14.3
'Compare'	84	0.2
Synonym description	310	0.8
	38,558	

Table 1 Source of extracted synonyms

Explicit references account for 898 synonym entries, and represent only a small percentage of the dictionary's potential (see table 1). Another source of synonyms are the fields flagged with **also** (e.g. *Abomasum[1]: The fourth or true digestive stomach of a ruminant:* ***also called reed****. Also ab.o.ma.sus*). The example shows that care must be taken to distinguish those cases where **also** introduces a spelling variant, which can be considered a synonym only in a very specialized sense. The **also** field contributes 2.4 % of the total synonyms extracted (table 1).

Another minor source (84 instances) is the **compare** field, aimed at the human dictionary browser. It points to words that are not quite synonyms, but that are semantically related (e.g. *Collage[1]: n. An artistic composition consisting of or including flat materials pasted on a picture surface; ···* ***Compare ASSEMBLAGE (def. 4)***).

A unique feature of the *F&W* dictionary are its **collateral adjectives**, which yield a small (0.3 %) but important set of 'synonyms'. They are flagged by ♦ and indicate adjectival forms of a noun entry "so remote in spelling that they may not be brought to mind by the noun" (preface, p 7a) (e.g. *Arm[1]: n. 1. Anat. An upper limb of the human body, from the shoulder to the hand or wrist.* ♦ ***Collateral adjective: brachial***). Their importance comes from the fact that the synonym crosses part-of-speech (POS) boundaries while remaining closely semantically related to headword.

INCREASE COVERAGE

From the above discussion it is clear that the dictionary has a quasi-formal structure with specific fields for each entry, which makes the extraction of information from it easier than if it were unrestricted free-format text, as for example a machine readable technical textbook. Building a grammar and writing a parser to convert the machine readable dictionary into a computerized dictionary is no trivial task, however, as many of the clues to interpret its structure are designed with a human reader in mind (the visual layout, for example, is important). The task of bracketing the fields in the *F&W* dictionary was made even more difficult by the fact that most of the typographical clues (italics, bold, superscript, etc) were lost in the process of making it machine readable: using a scanner and optical-character-recognition software, the book was stored in plain ascii format (to make matters worse, the scanning process introduced some errors of its own, especially damaging when they involve critical flags like sense numbers; e.g. the letter 'l' in stead of number 1, and alphabetic 'Oh' rather than zero in 1O. See WILMS 90).

If only the above relations were used, all of which are explicitly flagged in the dictionary, the extracted lists would compare very poorly with the wealth of information found in a thesaurus. However, many more can be found by looking at the actual definitions of each lemma. Definitions in the *F&W*, like most dictionaries, follow the Aristotelian principle of *genus* (supertype) and *differentiae* (necessary and sufficient conditions that separate it). A working hypothesis adopted in this paper is that definitions consisting only of a *genus* can be treated as synonyms. Practically, this means that **one-word definitions** can be treated as a synonym. And whereas one-word definitions are bad lexicographic practice that *F&W* claims never to be guilty of (preface, p. 6a), many definitions are actually multi-part explanations, some portions of which are single words (e.g. *Call[1]: v.t. 1. To say in a loud voice;* ***shout****;* ***proclaim****. 2. To* ***summon****. 3. To* ***convoke****;* ***convene****: to call a meeting. 4. To invoke solemnly*). Notice that sense number two is an example of something *F&W* claims never to do. Some liberty is taken in specifying single-word definitions: fillers like *a(n), the, to, any* are removed to leave true single genuses.

Thus at the small cost of some extra parsing overhead, the number of synonyms is enormously increased (30,720 entries, or 79.7% of the total; see table 1). The following strategies, with varying degrees of parsing sophistication, also contribute to increase coverage:

1. Ignore differentiae that are not so 'necessary and sufficient', and contribute little. For instance, phrases introduced by *as*; their purpose is to give a (prototypical) example that is not meant to be exclusive (e.g. *Collect[1]: ··· 6. To accumulate,* ***as sand or dust***). Sometimes the purpose of

including them seems to be a justification on the part of lexicographer for isolating and adding a separate sense (e.g. *Cahoots[1]: n. pl U.S. Slang Affiliation; partnership, as in the phrase* ***in cahoots***). Similarly, *parenthesized expressions* sometimes are illustrative objects (e.g. *Abstract[1]: v.t. ... 3. To withdraw or disengage* ***(the attention, interest, etc.)***), extra (non-'necessary') information (e.g. *A[1]: ... 4. Chem. Argon* ***(symbol A)***) or optional elements (e.g. *Absent[1]: v.t. To take or keep* ***(oneself)*** *away*).

2. **Relax the single-word criterion**. By *F&W* own standards the treatment of synonymy includes 'phrases and circumlocutions' (see the quote from the preface above). Thus idioms (compounds whose meaning is more or different than the definition of its parts) may be treated as a unit (e.g. *Acolyte[1]: n. 1. An attendant or assistant. 2. An* ***altar boy*** or *Anthracite[1]: n. Coal that burns slowly and with great heat: also called* ***hard coal***). This also includes idiomatic verb-preposition compounds (e.g. *Repress[1]: v.t. 1. To keep under restraint or control. 2. To* ***cut down****; quell, as a rebellion*). The necessary information to identify such compounds is supplied by the dictionary itself! During an initial bootstrapping pass, a list was generated that includes all headwords (both simple lemmas and compound entities, e.g. ***State****[1]: n. 1. Mode of existence as determined by circumstances, external or internal; nature; condition; situation. 2. Frame of mind; mood. 3. Mode or style of living; station. ... — to lie in state To be placed on public view, with ceremony and honors, before burial. — v.t. stat.ed, stat.ing 1. To set forth explicitly in speech or writing; assert; declare. 2. To fix; determine; settle. — sta.tal adj* and ***State policeman****[1]: U.S. A member of the separate police force of a State; also called State trooper, trooper*), their conjugated forms (e.g. *stated, stating*), derivatives (e.g. *statal*), and expressions (e.g. *lie in state*), with their POS (part-of-speech). This information will also be exploited to disambiguate or-conjoined phrases (see below). 1443 or 3.7 % of the extracted synonyms are compounds.

3. Increase the sophistication of the parser. There is a certain point of diminishing returns; extracting additional information is possible only at the cost of increasing parsing difficulty. Whereas the general layout of a dictionary entry is moderately structured, with fairly recognizable and limited sets of field separators, the content of those fields is much less restricted. Especially the defining text is hard to parse, and often consists of ellipses and incomplete phrases. In a way this is a chicken-or-egg problem: researchers turn to machine-readable dictionaries for lexical information for their natural language processing programs, but one such program, a wide-coverage parser, is needed to access much of the information buried in the dictionary.

One way around this is to **use heuristics**. It is not much more difficult to recognize two conjoined 'one-word' definitions than it is to find a single one (e.g. *Abhorrent[1]: adj. 1.* ***Repugnant or detestable***). Even detecting a list of multiple candidates (e.g. *Allied[1]: adj. 1.* ***United, confederated, or leagued***) is fairly fool-proof. But if one side of the conjunction consists of two (or more) words (templates like *X or Y Z* and *Z Y or X*, assuming *Y Z* or *Z Y* isn't a compound), this can cause problems for synonym candidate *X* because *Z* may have to be distributed, in which case we are no longer dealing with a 'single' word (i.e. the templates may be equivalent to *X* ***Z*** *or Y Z* and *Z Y or* ***Z*** *X*, respectively). If *X* is a *noun* on the *left* side of the conjunction (template *X or Y Z*), it is safe in most situations to accept it as a synonym (e.g. *Abeyance[1]: n. 1.* ***suspension*** *or temporary inaction*). But when the noun appears on the *right* hand side (template *Z Y or X*), it is often unclear even to native speakers whether *Z* should be distributed (shared by both *Y* and *X*. E.g. *Althorn[1]: n. An alto flügelhorn or* ***[alto???]*** *saxhorn*). Humans can often disambiguate these cases by relying on world knowledge, for which unfortunately there is no 'heuristic' (e.g. *Amerindian[1]: n. An* ***American*** *Indian or Eskimo*). For *verbs* the situation is reversed, and distribution is unclear for the verb on the *left* (e.g. *Adulate[1]: v.t. To flatter* ***[extravagantly???]*** *or praise extravagantly*), but unambiguous for verbs on the right (e.g. *Abandon[1]: v.t. 1. To give up wholly; desert; forsake. 2. To give over or* ***surrender****: with to*). Heuristics like pattern matching may offer a way out, though often it seems easier to rule out a candidate than to confirm one (but surprisingly candidate-selection rules extract many more synonyms than candidate-rejection rules (37,246 as compared to 36,332)). One simple rule that has great disambiguating power states that a synonym candidate must have the same POS as the head-word (POS information, as mentioned earlier, comes from using the dictionary source as a resource). This rules out cases like *Abhorrence[1]: n. 1. A feeling of utter loathing. 2. something loathsome or* ***repugnant*** and *Adventure[1]: n. 1.* ***hazardous*** *or perilous undertaking*. The matching-POS heuristic fails to disambiguate, however, when no speech information is available or may make wrong decisions when the candidate has multiple POS (e.g. the heuristic fails to rule out the following candidate, because *visionary* happens to be a noun in addition to being an adjective: *Abstraction[1]: n. ... 3. A* ***visionary*** *or impractical theory*. At the same time, there are instances when the algorithm undergenerates, as in *Adherent[1]: adj.* ***Clinging*** *or sticking fast*, because there is no entry for *clinging* in *F&W* (except for being mentioned as a gerund in the entry for *cling*)). Thus when the synonym candidate *X* does pass the POS test, other checks are necessary to cope with possible distribution problems.

4. **Explore Extended Differentiae**. Another unique feature of the *F&W* dictionary is the presence of occasional exposés that make a point about grammar or semantics. Written in an informal style, these asides seem like an opportunity for the lexicographer to address the reader directly and make something clear, usually

involving a subtle difference in usage between several lemmas (e.g. *ain't - aren't, any one - anyone*, etc.). Since it is debatable whether 'true' synonyms exist that are interchangeable in *all* contexts, many of the one-word definitions in the dictionary lack differentiae only because these differentiae are too subtle to allow explaining in a short sentence. These mini-essays are an attempt by the lexicographer to provide some more context and examples to quantify some of the differences in usage. In many cases the program has no difficulty in identifying the synonyms in the text (e.g. *Adherent[1]: … — (noun)* ***Adherent*** *is the weakest term. A* ***follower*** *is more fervid in his attachment. A* ***disciple*** *has a pupil-teacher relationship with the one he follows. A* ***supporter*** *is one who aids in any way, while a* ***partisan*** *is militant in his support.*). It is dangerous to assume, however, that the first word is a synonym. As a precaution, the programs verifies that the candidate has the correct POS. The combination of selectiveness and uneven exhaustiveness of a dictionary can be both an advantage (e.g. no entry for *Rec't*, so the heuristic correctly rejects the first word as a synonym candidate: *Abbreviation[1]: … — Syn. 1. An abbreviation is a shortening by any method. A contraction is made by omitting certain medial elements (whether sounds or letters) and bringing together the first and last elements.* ***Rec't*** *for receipt is a written contraction as well as abbreviation*) and a disadvantage (e.g. *man* happens to also be an adjective and thus seems to qualify in *Adequate[1]: … — Syn. 1. Adequate is applied to ability or power; sufficient, to quantity or number. A* ***man*** *is adequate to a situation*). In some cases additional heuristics can help to detect inappropriate matches; the presence of a previously accepted synonym in the middle of a sentence establishes that sentence as an illustrative example rather than the introduction of a new synonym. Examples can sometimes be used as additional proof to confirm a candidate, or as evidence against an inappropriate choice (e.g. the tentative selection of *both* (which is also an adjective) can be overturned because it is not listed among the examples: *Addicted[1]: … — Syn. Addicted suggests a pathological weakness; given, a tendency or usual practice.* ***Both*** *words may apply to good or bad things, but usually to bad: addicted to alcohol, given to lying*). Since the program does not perform any semantic analysis but relies instead on syntactic clues, there will always be some under- and over-generation (e.g. *Acumen[1]: n. … — Syn. 1. Sharpness, acuteness, and insight, however keen, and* ***perception****, however deep, fall short of the meaning of* ***acumen****, which belongs to an astute and discriminating mind*).

SENSE (PLUS LEMMA NUMBER AND POS) DISAMBIGUATION

From the above examples it can be seen that each lemma in the dictionary has three features: lemma number, POS, and sense number. The **lemma number** (indicated in superscript) distinguishes words that 'accidentally' share identical spellings but are of quite different origin (e.g. *Prune[1]: n. The dried fruit of the plum.* ***[< OF < LL < L prunum]*** and *Prune[2]: v.t. & v.i. … To cut off (branches or parts).* ***[< OF proöignier, proignier, ? < provaignier to cut]***); there are 1058 instances of these multiple-entry lemmas in the *F&W* dictionary. Only in a few minor instances (12 to be exact) do the lexicographers explicitly indicate the lemma number for the 'synonym' (e.g. *Chine[2]: n. Chime[2]*); these exceptions are all true one-word definitions. It is understandable that a dictionary does not indicate the lemma number for every word in the definition text; therefore the program considers the lemma number of the extracted synonym as an unknown to be disambiguated. By treating these multiple-entry lemmas as completely unrelated entries, the program can avoid the pitfall of relating synonyms on the basis of orthographic rather than semantic similarity (e.g. *tough, unfeeling* and *hard* are related to *Rocky[1]: adj. … Consisting of, abounding in, or resembling rocks*, but have nothing in common with *Rocky[2]: adj. … Inclined to rock or shake*: *unsteady, dizzy, weak*).

Within each dictionary entry, lemmas are further grouped by **part-of-speech** (POS, flagged by a pair of *[]*). There are 17 classes of POS in *F&W*, and only *auxiliary* does not have any synonyms (see table 2). With a few exceptions all synonyms inherit the POS information of their headword (for *prefix, symbol*, and *combining form* that is usually not the case, and thus the field is left open, to be filled out as part of the disambiguation process; e.g. *Supra-[1]* ***prefix****. Above; beyond*). Idiomatic expressions and compounds are listed in the *F&W* dictionary at the end of the main entry block, but without POS information (and the grammatical category of the expression many times is different from that of the headword; e.g. *Head[1]: n. …* ***— head over heels*** *1. End over end. 2.Rashly; impetuously. 3. Entirely; totally. —* ***to have a head*** *Informal To have a bad headache. —* ***to make head or tail of*** *To understand: usu. used in the negative*). In many cases an expression can at least be identified as verbal *[v]* (or

	Distribution of Headwords Total: 27,103	Distribution of Synonyms Total: 38,558	Average number of Synonyms per Headword	Percent of total number of Synonyms
Noun	9,351	12,550	1.34	32.5
Adjective	7,034	10,815	1.54	28.0
Verb Transitive	4,821	7,019	1.46	18.2
None or Unknown	1,300	2,789		7.2
Verb Intransitive	1,890	2,582	1.37	6.7
Verb	958	1,460	1.52	3.8
Adverb	699	969	1.39	2.5
Combining Form	526	0		0
preposition	146	184	1.26	0.5
Prefix	118	0		0
Suffix	86	0		0
Conjunction	67	83	1.24	0.2
Interjection	52	66	1.27	0.2
Symbol	29	0		0
Pronoun	24	36	1.50	0.1
Indefinite Article	1	4	4.00	0
Pronominal Adjective	1	1	1.00	0

Table 2 POS distribution of the synonyms

even more specifically transitive *[vt]* or intransitive *[vi]* for verbs that aren't both) because of the leading *to* when followed by a verb (e.g. in the above *head* example, *to* ***have*** *a head* is assigned *[vt]*; *to* ***make*** *head or tail of* and hence its synonym *understand* are assigned generic *[v]*). But even this heuristic is not fool-proof, again because a dictionary also includes peripheral senses and POS (e.g. the program would incorrectly assign *[vt]* to the expression: *Advantage[1]: n. ··· — **to advantage** To good effect. — v.t. To give advantage or profit to*). During the disambiguation phase the program will attempt to resolve any synonyms which inherited a generic *[v]* POS to the more specific *[vt]* or *[vi]* (see below).

Finally, for each POS a lemma may have **multiple senses** (flagged by a pair of <>). As with lemma numbers, it is important to distinguish between the different meanings of a lemma (e.g. *comical* and *humorous* are synonyms of *Funny[1] [adj]<**1**>*, but should not be included in the set *peculiar*, *strange*, *odd* which are listed under *Funny[1] [adj]<**2**>*). Unfortunately, explicit sense references (e.g. *Portion[1]: n. ··· 5. A dowry **(def. 1)***) are rather rare (146 synonyms, or 0.4 %), and usually refer to a peripheral or field-specific (theater, physics,...) sense of the synonym. Four of these instances are unusual in the sense that both members of the synonym pair point to *different* senses of each other (e.g. *Pup[1]: n. 1. A **puppy** (def. **1**). 2. The young of the seal, the shark, and certain other animals* and *Puppy[1]: n. 1. A young dog: also called pup. 2. A **pup** (def. **2**)*).

For most **headwords,** the value of these three fields (lemma number, POS, and sense number) is obviously explicit, with two exceptions: the POS for compounds is omitted, even for those that have an entry of their own (see above); more critical are the instances where the dictionary does not specify the appropriate sense of a headword when listing synonyms! This occurs 370 times for explicit references, and 898 times for synonym descriptions. In the case of *explicit references*, the missing sense number seems to be an oversight on the part of the lexicographer, rather than an indication that the synonym applies to all senses of the headword (e.g. *Given[1]: adj. 1. Presented; bestowed. **2. Habitually inclined;** 3. Specified; stated: a given date. 4. Issued on an indicated date: said of official documents, etc. 5. Admitted as a fact. — **Syn. See ADDICTED***). The oversight hypothesis is supported by the fact that in some cases where the synonym does map onto several senses of the headword, the numbers *are* given (e.g. *Harmony[1]: n. 1. Accord or agreement in feeling, manner, action, etc. 2. A state of order, agreement, or esthetically pleasing relationships among the elements of a whole. 3. ··· — Syn. **1, 2.** Concord, accord, consonance, congruity*). With some *synonym descriptions*, however, the lack of a sense number truly reflects that the synonyms refer to several senses (e.g. *Calm[1]: adj. 1. Free from agitation; still or nearly still. 2. Not excited by passion or emotion; peaceful. — Syn. (adj.) ··· A **placid person** is regarded as temperamentally stolid; a **placid lake** is always peaceful*), or that some of the synonyms in the group refer to one sense of the headword, whereas others map onto a different sense (e.g. *Expel[1]: v.t. 1. To drive out by force. 2. To force to end attendance at a school, terminate membership, etc.: oust. — Syn. A school **expels** an unruly **pupil**; **water** in the lungs must be promptly **expelled***).

	Explicit POS		Unknown POS		
	Explicit sense number	Unknown sense number	Explicit sense number	Unknown sense number	
Explicit lemma number	1	6	1	4	12
Unknown lemma number	114	34,188	30	4,214	38,546
		34,309		4,249	

Table 3 Distribution of bound/uninstantiated field variables for synonyms

	_ ? # _		_ ? m _		_ ? * _	
	___?	___x	___?	___x	___?	___x
#___	756		535		25	
x___	535	26	1952	20	151	
*___	25		151		66	

Table 4 "Legal" *HS?-S'H'* combinations for disambiguating *S*

Table 3 shows the distribution of the three potential unknowns for the **synonyms** (lemma number, POS, and sense number). The only case in the whole *F&W* where all three variables are instantiated is *Scansion*[1] (*n. The division or analysis of lines of verse according to a metrical pattern. Compare* ***METER*[2] *(def. 1)***), and it is not symmetric (i.e. *Meter*[2] does not list *Scansion* as a synonym)!

One approach for performing sense-disambiguation suggested by Chodorow is **disambiguation by symmetry** (CHODOROW 88); given a headword *H* whose *x*th sense refers to a synonym *S*, then to find what sense number *y* of synonym *S* is the appropriate match, scan the definition text of all lemmas *S'* for a reference back to *H'* (this strategy will have the fringe benefit of also disambiguating the lemma number and POS of *S* if they are unknown). For example, the *second* sense of *differ* is synonym with *disagree*[1] (*v.i.* ***1****. To vary in opinion;* ***differ****; dissent. 2. To quarrel; argue*) by virtue of circularity (*v.i. 1. To be unlike in quality, degree, etc.: often with from.* ***2****. To* ***disagree****: often with with. 3. To quarrel*).

Assume *?* to stand for 'unknown', *x y* to be digits representing sense numbers, *#* to mean there is only one sense, and * to indicate the headword *H* is not sense-disambiguated (see above); then in pairing *HS* and *S'H'* to sense-disambiguate *S*, there are theoretically 36 combinations possible:

the sense number of *H* is either known (*#* if only one, *x* if polysemous) or unspecified (*)
the sense number of *S* is either explicit (*y*) or uninstantiated (*?*)
the sense number of *S'*, like *H*, is ordinarily known (*#* or *y*), or unspecified (*)
the sense number of *H'*, like *S*, is either given (*x*) or unknown (*?*).

Actually, 11 of these potential combinations are either '*wrong*' or '*redundant*': *H*<_>*S*<*y*>-*S'*<#>*H'*<_> (and matching *H*<#>*S*<_>-*S'*<_>*H'*<*x*>) i.e. no matter what the values are for *H* and *H'*, if *S'* has but one sense it is wrong to list a number for *S* (or at best redundant if *y*=1, especially considering the fact that explicit sense numbers occur so rarely). There is one such '*error*' in *F&W*, probably attributable to the fact that composing a dictionary is a group effort that takes many years (i.e. while working at the letter *D* it is difficult to anticipate the details of the *S* volume). Compare *Davy*[1]*: n. A safety lamp* ***(def. 1)*** *with Safety lamp*[1]*: A miner's lamp having the flame surrounded by fine wire gauze that prevents the ignition of explosive gases: also called davy*.

Of the remaining 25 '*legal*' combinations, 13 do actually occur in *F&W*: there are 4,290 matches (2,145 *symmetric* synonym pairs). The majority of them participate in the sense disambiguation process, i.e. match the pattern *H*<_>*S*<**?**>-*S'*<_>*H'*<_> (see table 4). In 46 cases the synonym pair is already sense disambiguated in one direction (e.g. *Lullaby (def. 1)* is a synonym of *Cradlesong*[1], but the dictionary does not specify what sense of *Cradlesong* matches with *Lullaby*[1]); in most cases the sense number must be disambiguated in both directions. In **column one and two** of table 4, the unknown sense of the synonym (*?*) is instantiated with 100 % confidence to **#**, i.e. there is no contest as the synonym has only one

sense for the POS in question. Notice that the second most common pattern belongs to this category: a headword with only one sense teaming up with a synonym with a single sense.

By far the most frequent synonym pairs, however, are those where both words are polysemous. Finding a sense *y* that points back to the original headword is in most cases conclusive evidence to replace *?* with *y* (**columns three and four**), but because there is always the possibility that another sense *z* is a more appropriate match, the confidence factor is slightly lower than in the previous (one-sense only) category.

Finally, in **columns five and six** of table 4 are those cases where *?* is instantiated to *****, i.e. circularity is established, but the problem of what sense of *S* matches with *H* has not been resolved yet. The program has an arsenal of four strategies, in decreasing order of confidence, to attempt to select the right sense. However, even if none of them succeed in sense-disambiguating *S*, having established circularity is sufficient to interpret the two words as being semantically closer than if the synonym relation had only been one-way.

1. **There is multiple evidence of circularity**, and the other proof is more specific. E.g. *Fragile<#>Frail<?>* maps with both *Frail<2>Fragile<?>* and with the generic *Frail<*>Fragile<?>* (*Frail[1]: adj. 1. Delicately constituted; weak. 2.* ***Fragile****. 3. Deficient in moral strength. — Syn. See* ***FRAGILE***). This approach works for 24 % of the * cases.

2. **Chodorow strategy number 2**: look at the **intersection** of the definitions of both words, i.e. do they **share** any **third synonym**? (CHODOROW 88) *Comical<#>-Humorous<?>* can be resolved to sense 1 because that sense shares *funny* as a synonym with *comical* (*Comical[1]: adj. Causing merriment;* ***funny****; ludicrous. — Syn. See* ***HUMOROUS***. and *Humorous[1]: adj.* ***1****. Full of or characterized by humor; laughable;* ***funny****. 2. Displaying or using humor. — Syn.* ***Comical****, droll, witty*).

3. **Try relaxing the working definition of synonymy**, i.e. establish multiple evidence (cf. approach 1) by accepting genuses with **non-empty differentiae** as synonyms; e.g. *Destroy<2>Demolish<?>* gets instantiated to sense 2 (*Demolish[1]: v.t. 1. To tear down, as a building. 2. To* ***destroy utterly****; ruin*). Practically, this means checking all senses of the synonym for the occurrence of the headword in the genus position. If there is no match, try the same for any third synonyms in the definition of the headword (e.g. *Audacity[1]: n. 1.* ***Boldness****; daring. — Syn. See TEMERITY* and *Temerity[1]: n. Venturesome or* ***foolish boldness****; rashness*).

4. Finally, as a last resort, relax the intersection approach for a **maximum overlap in function words**; This approach doesn't help if the headword consists of one-word definition(s), but then the definition text of a third synonym may help; e.g. *Ardent[1]* (*adj. 1.* ***Passionate****; zealous;* ***intense****. ...*) matches with the third sense of *Intense[1]* (*adj. 1. Having great force; overpowering: intense feelings. 2. Performed strenuously and steadily: intense study.* ***3. Expressing strong emotion****: an intense look; also,* ***characterized*** *by* ***strong*** *and earnest feelings: an intense person*) because that sense has a strong overlap with *Passionate[1]<3>* (*Expressing, displaying, or characterized by passion or strong emotion;* ***ardent***). The disambiguation resulting from this heuristic carries a low confidence weight, however, because even when insisting on intersections of at least size three, the risk is great of making incorrect assignments because of accidental, non-meaningful overlap.

ASYMMETRIC SYNONYMS

Chodorow reported that about 62 % of the synonyms in *the New Collins Thesaurus* are asymmetric (and thus cannot be sense-disambiguated using his first strategy) (CHODOROW 88). For the *F&W* that figure is much higher: 88 %. In the majority of the cases this means that while one of the senses of the synonym does match with the headword, there is no explicit pointer back to the headword. For a small subset (98 cases) of asymmetric synonyms there is actually an explicit sense pointer in the definition of the headword. For 25 % of the asymmetric synonyms, the sense-disambiguation problem can be very quickly solved by simply noticing that the synonym has only one sense for the POS in question (this works even for the cases with multiple lemma numbers; as long as none of those entries has more than one sense, the sense number can be disambiguated to #, even as the lemma number remains unresolved). For the others, any of the last three strategies outlined above may do the trick.

A fortunate side-effect of sense-disambiguation is that it also automatically assigns lemma numbers and resolves unknown POS (see above). For asymmetric synonyms, the same table-lookup routine that disambiguates senses on the basis of finding only one sense, also succeeds in binding 86 % of the unresolved lemma numbers (because lemmas with multiple lemma numbers are a minority). Similarly, 33 % of the generic *verb* POS can be narrowed down to *transitive* or *intransitive verb* (e.g. *Run[1]: ... — to run off ... 3. To flee or escape;* ***elope***. Since the last synonym can only be used intransitively, it can be disambiguated from *Run off[1][V]<3>Elope[2][V]<?>* to *Elope[1]**[VI]**<?>* despite the fact that it has no pointer back to *run off*). In the same manner 37% of the unknown POS cases become resolved to a more specific POS.

CONCLUSION AND FUTURE RESEARCH

6.5 % of all synonyms remain unresolved because they

have no separate entry in the *F&W* dictionary; a random sample of 50 pairs of unresolved asymmetric synonyms shows that 8 % are compound headwords that are only listed as derivatives of their main entry; also 4 % of the synonyms exist only as derivatives. In 10 % of the samples the synonym *did* actually point back to the headword, but errors in the dictionary (in the on-line version, *and* in the printed original) misled the parser, or the parser wasn't sophisticated enough to recognize some synonyms. Four percent of the asymmetric synonyms were incorrectly flagged as synonyms by the parser. Another 4 % of the synonyms are actually hypernyms, which are by nature asymmetric. 30 %, finally, of the non-circular synonyms involve headwords that are used in a peripheral sense (often with labels like *slang, informal, ...*). One of the future goals is to sophisticate both the parser and the disambiguator (especially the heuristic of overlapping function words) to reduce the number of these unresolved cases (Both programs are written in Pascal, and run on a Sun Spark Station; the dictionary source occupies about 9 Meg of disk space).

Another possible direction of research is to investigate whether the hypothesis of Veronis et al. that "it is extremely unlikely that the same information is consistently missing from all dictionaries" also holds for synonyms, as about 20 % of the asymmetric synonyms seem to be inconsistencies and oversights on the part of the lexicographers (VERONIS 90, p. 231). In this case merging several machine readable dictionaries may correct many of these omissions.

Once a high quality pool of sense-disambiguated synonyms has been generated, the next exploration will be to expand this list by traversing chains of synonyms and to generate a taxonomy of related words with distance weights.

SELECTED BIBLIOGRAPHY

ATKINS B, LEVIN B (1991) Admitting Impediments. *Lexical Acquisition: Exploiting Online Resources to Build a Lexicon* edited by ZERNIK U. Lawrence Erlbaum Associates Inc. NJ, pp 233-262.

AMSLER R (1981) A Taxonomy for English Nouns and Verbs. Proceedings of the 19th *Annual Meeting of ACL*, pp 133-138.

BYRD R (1983) Word Formation in NLP Systems. *IJCAI 83 (8th): Germany*, pp 704-706.

CALZOLARI N (1984) Detecting Patterns in a Lexical Data Base. *COLING 84 (10th) (22nd Meeting of ACL)*, Stanford, CA, pp 170-173.

CHODOROW M, RAVIN Y, SACHAR H (1988) A Tool for Investigating the Synonymy Relation in a Sense Disambiguated Thesaurus. Proceedings of the 2nd *Conference on Applied NLP: ACL*, Texas, pp 144-151.

HODGES J, BOGGESS L, CORDOVA J, AGARWAL R, DAVIS R (1991) The Automated Building and Updating of a Knowledgebase through the Analysis of Natural Language Text. *Technical Report MSU-910918*: Mississippi State University

KAZMAN R (1986) *Structuring the Text of the OED through Finite State Transduction*. Waterloo, Ontario.

OSGOOD C, SUCI G, TANNENBAUM P (1971) *The Measurement of Meaning*. University of Illinois Press, Urbana.

SPARCK JONES K (1964) *Synonymy and Semantic Classification*. PH D Thesis, University of Cambridge, England.

VERONIS J, IDE N (1990) Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. Proceedings of 13th *COLING 90*, Helsinki, pp 389-394.

VERONIS J, IDE N (1991) An Assessment of Semantic Information Automatically Extracted from Machine Readable Dictionaries. Proceedings of 5th *EACL Conference*, Berlin, Germany, pp 227-232.

WILMS G (1990) Computerizing a Machine Readable Dictionary. Proceedings of 28th *ACM Southeast Region Conference*, South Carolina, pp 306-313.