Tools for Lexicographers Revising an On-Line Thesaurus

Yael Ravin, Martin S. Chodorow, and Howard E. Sachar

Introduction

Revision and extension of published dictionaries and thesauri form an essential component of the work of lexicographers. There are inherent difficulties associated with these tasks due to the large volume of data involved: consistency is hard to maintain, and checking or testing can become extremely tedious. These difficulties may be substantially aided by computer programs which manipulate the data, sort them in various ways, and present different relevant portions of the text to the lexicographer, to make decisions and instantiate changes. This paper discusses the automatic data manipulation that we perform as part of our lexical work at the IBM Watson Research Center and the ways in which it is relevant to lexicographers.

Our research interest is in equipping the computer with lexical knowledge. The authors' recent efforts (Chodorow et al. 1988) have concentrated on equipping the system with some knowledge of synonyms derived from the machine-readable version of THE NEW COLLINS THESAURUS (henceforth CT).[1] Unlike humans, computers cannot rely on their "common sense", so information that is implied or assumed in CT had to be made explicit. For example, headwords had to be supplied with their parts of speech, and synonyms had to be disambiguated.[2] Because of the size of the source, these tasks had to be performed automatically.

In our computational manipulation of the CT material, we discovered some interesting properties of the interconnections found in the thesaurus: many of the links between synonyms are asymmetric and many are intransitive. These properties of asymmetry and intransitivity are common to most thesauri but their extent differs according to the size of the book and the judgements made by its lexicographers.[3] Thus, the individual character of a particular thesaurus and its lexical content can be captured by a description of the patterns of asymmetry and intransitivity found in it. Moreover, asymmetry and intransitivity are the product of human judgement, in situations often involving conflicting criteria. Consequently, inconsistency may very likely exist in the finished product. In the process of lexicographic revision, listings of asymmetry and intransitivity would seem useful.

In the first section of this paper, we describe asymmetry and intransitivity as they appear in CT and discuss the concept of synonymy they express in the book. In the second section of the paper we describe how we have automatically disambiguated the synonyms found in CT. We had to perform sense disambiguation in order to be able to refer to particular senses of words, because synonymy links exist between *senses* of words, not between words themselves. In the rest of the paper, we discuss asymmetry (and, briefly, intransitivity) and suggest ways in which it can be captured and corrected.

Properties of CT-synonyms

WEBSTER'S SEVENTH NEW COLLEGIATE DICTIONARY (henceforth W7) defines a thesaurus, such as CT, as "a book of words and their synonyms". But what are synonyms? The definition and existence of synonyms have long been debated in linguistics. Some believe it is impossible to capture meaning, not even of the most concrete terms. Consequently, it is impossible to define synonymy or to identify synonymous terms (Quine 1960). Others believe it is possible to give full semantic representations of meaning and therefore to define synonymy formally and to identify true synonyms (Katz and Fodor 1963). According to this view, synonymy is a relationship of sameness of meaning between words, which is defined as the identity of their semantic representations. We have chosen an operational approach to synonymy: The synonyms of a headword *w* are whatever words are listed in the entry for *w* in the on-line version of CT. According to the authors, "...no synonym is entered unless it is *fully* substitutable for the headword in a sensible English sentence" (CT 1984:v). This may suggest that each entry (i.e. a headword and its synonym list) contains all and only words that are closely related semantically. But the same synonyms appear in several lists, and headwords are themselves synonyms of other headwords, so that the lists in CT are implicitly interconnected.

The links in the thesaurus can be characterized according to their degree of symmetry and transitivity. We say that the link between *a* and *b* is *symmetric* if *a* points to *b* and *b* points to *a*; that is, if the headword *a* has *b* in its synonym list and the headword *b* has *a* in its list. We say that the link between *a* and *b* is *transitive* if for every word *c*, if *b* points to it then *a* points to it too; that is, if all the synonyms found in *a*'s synonym list are also found in *b*'s list (with the exception of *a* and *b* themselves, of course). Thus, if links were symmetric and transitive throughout the thesaurus, all words would partition into disjoint sets. Each member of the set would be a synonym of every other member.

There are only 27 sets of words in CT which exhibit completely symmetric and transitive links among their members. Within the context of the thesaurus, these may be considered to have identical meaning. 26 out of the 27 are word pairs — the 27th is a triple — and all have a single sense and a unique part of speech.[4] These sets are given below.

allocate	=	allot
aphorism	=	apothegm
astonishing	=	astounding
at_times	=	from_time_to_time
bystander	=	eyewitness
cemetery	=	necropolis
congratulate	=	felicitate
eatable	=	edible
entomb	=	inter
everybody	=	everyone
exactitude	=	exactness
greetings	=	regards
insomnia	=	sleeplessness
lozenge	=	pastille
myopic	=	near-sighted

naught	=	nought
perk	=	perquisite
permeable	=	porous
piddling	=	piffling
podium	=	rostrum
prizefighter	=	pugilist
prizefighting	=	pugilism
saw	=	saying
slattern	=	slut
testy	=	tetchy
triad	=	trinity = trio
weal	=	welt

Most of the synonymy links in CT are markedly different from these. 62% are asymmetric (e.g., *part* has *department* as a synonym, but *department* does not have *part*); and 65% are non-transitive (e.g., *part* has *piece* as a synonym; *piece* has *chunk* as a synonym; but *part* does not have *chunk* as a synonym).[5]

According to the substitutability definition of synonymy adopted by Collins, links should always be symmetric since if it is possible to substitute *b* for *a* in a "sensible" English context, then it is always possible to reintroduce *a* into that context as a substitution for *b*. For similar reasons, links should always be transitive. On the other hand, lack of symmetry and transitivity may be purposely chosen by the lexicographer because of other considerations that are involved, such as usefulness to a human reader, constraints on space, and aesthetic presentation. These considerations often override the substitutability criterion to result in some asymmetry and intransitivity. The particular resolution of this conflict for each entry gives the thesaurus its individual character. It is also a potential source of inconsistencies, which can be revealed by automatic means. In fact, we collected much of our data while attempting to automatically disambiguate the senses of CT-synonyms.

Sense Disambiguation

Since synonymy links occur between senses of words and not between words themselves, we found it necessary to disambiguate the words given in the CT synonym lists, so that we would be able to refer to a particular sense of each synonym.

Every entry in CT is broken into the different senses of its headword, as can be seen in the entry of *house*, given below, which contains 6 senses.

1. abode, building, domicile, dwelling, edifice, habitation, home, homestead, residence
2. family, household, ménage
3. ancestry, clan, dynasty, family tree, kindred, line, lineage, race, tribe
4. business, company, concern, establishment, firm, organization, outfit (Informal), partnership
5. Commons, legislative body, parliament
6. hotel, inn, public house, tavern

The synonyms listed for each sense, however, are not marked for their intended sense. Thus, it is not explicitly marked which sense of *abode*, for example, is linked to *house1*. We have tried two automatic methods of sense marking (i.e. sense disambiguation): disambiguation by symmetry and disambiguation by intersection.

In a dictionary-style thesaurus such as CT, an entry *a* may have word *b* listed as a synonym of its nth sense, and entry *b* may have word *a* listed as a synonym of its mth sense. We can mark *b* in entry *a* as the mth sense of *b*, and *a* in entry *b* as the nth sense of *a*. An example of this type of one-to-one mapping in CT is given below.

dense (adj)	1. ... condensed ... solid ....
	2. ... dull ... stupid ...
dull (adj)	1. dense .... stupid ....
	2. ... callous ... unsympathetic
	.
	.
	.
	7. drab ... muted ....

Here, sense 1 of *dull* is synonymous with sense 2 of *dense*. 37% of the 287,000 synonym tokens show this type of symmetry. Of course, there are also mappings of the one-to-many variety (for example, only the first sense of *feeble* has *faint* as its synonym, whereas both senses 1 and 2 of *faint* have *feeble*), but they account for only .5% of the tokens. By this method of disambiguation-by-symmetry, we could automatically mark the senses of all synonyms in one-to-one and one-to-many relations. The third type of mapping, many-to-many, accounts for just .5% of the total, but it poses a problem for the strategy outlined above. This can best be seen by considering an example. Senses 1 and 2 of *institution* list *establishment* as a synonym, and senses 1 and 2 of *establishment* list *institution*. Is sense 1 of *institution* synonymous with sense 1 of *establishment* or with sense 2? The distribution of the terms *institution* and *establishment* cannot answer the question.

The problem of many-to-many mappings and the large percentage of asymmetric CT-synonyms led us to another method. Consider again the case of *dense* and *dull*. Evidence for linking sense 2 of *dense* with sense 1 of *dull* comes from the symmetric distribution of the two words in the entries. There is however another piece of evidence for linking sense 2 of *dense* with sense 1 of *dull*, and that is the co-occurrence of the word *stupid* in their synonym lists. Thus, the intersections of synonym lists serve as the basis for an automatic disambiguation of the many-to-many mappings, and, for that matter, for the disambiguation of the whole CT. This is similar to Lesk's suggestion for disambiguating words in context (Lesk 1986). The intersection method disambiguated more entries than the symmetry method, but it, too, left a certain percentage of ambiguous words. In some cases, the intersection of two words was null. For example: *successful* and *victorious* are symmetric synonyms but none of their other synonyms are shared. Their entries are given below.[6]

SUCCESSFUL:
> > 0 acknowledged$ at_the_top_of_the_tree$99
best-selling$99 booming$99 efficacious$

favourable$ flourishing$0 fortunate$1.2
fruitful$3 lucky$1 lucrative$0
moneymaking$0 out_in_front$99 paying$99
profitable$1 prosperous$1 rewarding$0
thriving$0 top$ unbeaten$1 victorious$
wealthy$0

VICTORIOUS:
> > 0 champion$ conquering$99 first$
prizewinning$99 successful$
triumphant$0 vanquishing$99 winning$2

In other cases, there was a tie. For example, *ripe*2 has equal-size intersections with both *perfect*1 and *perfect*4. In their following entries, ties are indicated by a pair of numbers joined by a period.

PERFECT:
> > 1 absolute$1 complete$1.3 completed$99
consummate$2 entire$1.3 finished$2 full$1
out-and-out$ sheer$2 unadulterated$99
unalloyed$99 unmitigated$2 utter$99 whole$1
> > 4 accomplished$2 adept$1 experienced$1
expert$2 finished$1 masterly$0 polished$
practised$ skillful$0 skilled$0

RIPE:
> > 2 accomplished$1 complete$2 finished$
in_readiness$ perfect$1.4 prepared$1
ready$1

No disambiguation resulted in either of these cases. The results obtained with each method are shown in the following table:[7]

by symmetry:		
sense disambiguated:	103,648	(46.7%)
ties:	1,662	(0.7%)
remainder:	116,647	(52.5%)
Total number of synonyms available for processing:	221,957	
by intersection:		
sense disambiguated:	179,126	(80.7%)
ties:	6,029	(2.7%)
remainder:	36,802	(16.6%)
Total number of synonyms available for processing:	221,957	

Figure 1 Disambiguation Results

The quantitative advantage of the intersection method is evident. To determine the qualitative difference, we studied cases where the symmetry and the intersection methods conflicted. We compared fifty randomly selected entries. Of the approxim-

ately 900 synonyms listed in the entries, 337 were disambiguated by both methods. Of these, there were 33 pairs for which the two methods disagreed. 20 were symmetric ties, disambiguated by the intersection method. 5 were intersection ties, disambiguated by the symmetry method. The remaining 8 were given to two human reviewers. In 3 out of the 8, the reviewers could not determine which of the methods provided better disambiguation, as shown in the following example.

FEEBLE:

1. debilitated, delicate, doddering, effete, enervated, enfeebled, etiolated, exhausted, failing, faint, frail, infirm, languid, powerless, puny, shilpit (*Scottish), sickly, weak, weakened
2. flat, flimsy, inadequate, incompetent, indecisive, ineffective, ineffectual, inefficient, insignificant, insufficient, lame, paltry, poor, slight, tame, thin, unconvincing, weak

POOR:

1. badly off, broke (*Informal), destitute, hard up (*Informal), impecunious, impoverished, indigent, inneed, in want, necessitous, needly, on one's beam-ends, on one's uppers, on the rocks, penniless, penurious, poverty-stricken, skint (*BritishSlang), stony-broke (*BritishSlang)
2. deficient, exiguous, inadequate, incomplete,insufficient, lacking, meagre, miserable, niggardly, pitiable, reduced, scanty, skimpy, slight, sparse, straitened
3. below par, faulty, feeble, inferior, low-grade, mediocre, rotten (*Informal), rubbishy, second-rate, shabby, shoddy, sorry, substandard, unsatisfactory, valueless, weak, worthless
4. bad, bare, barren, depleted, exhausted, fruitless, impoverished, infertile, sterile, unfruitful, unproductive
5. hapless, ill-fated, luckless, miserable, pathetic, pitiable, unfortunate, unhappy, unlucky, wretched
6. humble, insignificant, lowly, mean, modest, paltry, plain, trivial

The symmetry method linked *feeble2* with *poor3*, whereas the intersection method linked *feeble2* with *poor2*. The remaining four cases were somewhat clearer. In three, the intersection method performed better; in one, the symmetry method was superior. To conclude, the best disambiguation algorithm would be a combination of the two methods. We are currently studying more cases where the methods disagree in order to determine how they should be combined.

Terminal Nodes

The largest source of asymmetry is *terminal* nodes: words that are offered as synonyms but do not occur as headwords. Thesauri typically contain terminal nodes as the number of synonyms usually exceeds the number of entries. In CT we found about 65,000 terminal nodes, accounting for 36% of the total of asymmetric links. 18,500 of them occur only once; but more than 400 occur 10 times or more. A sample of frequently occurring terminals is given below, with the number of their occurrences and a list of the entries in which they occur.

10 NONPLUSSED(adj): blank$3 confused$1 dazed$0 dumbfounded$0 dumfounded$0 flabbergasted$0 puzzled$0 stuck$2 surprised$0 thunderstruck$0

10 RECKLESSLY(adv): blindly$2 dangerously$1 fast$6 hastily$2 headfirst$2 helter-skelter$1 impetuously$0 incautiously$0 madly$3 pell-mell$1

10 PRIME MOVER(n): architect$2 author$0 cause$1 creator$0 father$3 instigator$0 mainspring$0 originator$0 prompter$2 protagonist$2

10 PROSECUTION(n): action$6 arraignment$0 enforcement$1 execution$1 furtherance$0 indictment$0 lawsuit$0 litigation$0 pursuance$0 suit$4

12 RIDGE(n): bank$2 bluff$3 crease$2 crest$1 knurl$0 ledge$0 projection$1 seam$3 wave$3 weal$0 welt$0 wheal$0

20 MITE(n): atom$0 bit$1 crumb$0 dot$1 dreg$0 grain$3 iota$0 jot$1 modicum$0 molecule$0 mote$0 particle$0 pennyworth$0 pinch$7 pittance$0 scrap$1 speck$2 tittle$0 tot$1 whit$0

23 RESTRICTED(adj): captive$2 cloistered$0 closed$3 cramped$1 dialectal$0 exclusive$2 exclusive$3 finite$0 hush-hush$0 incommodious$0 inside$5 light$25 limited$1 limited$2 local$2 narrow$1 parochial$0 peculiar$2 qualified$2 reserved$1 scanty$0 straitened$0 topical$2

28 REDUCTION(n): abasement$0 abatement$1 abbreviation$0 abridgment$0 alleviation$0 allowance$3 bargain$2 condensation$3 constriction$0 contraction$0 cut$9 cutback$0 debasement$1 decrease$2 deduction$2 depletion$0 diminution$0 discount$3 drain$6 drop$4 fall$11 lessening$0 retrenchment$0 saving$2 vitiation$1

28 RUDENESS(n): acerbity$1 audacity$2 awkwardness$1 brass$0 churlishness$0 contumely$0 crudity$2 discourtesy$1 disrespect$0 effrontery$0 grossness$2 impertinence$0 impoliteness$0 impudence$0 incivility$0 indelicacy$0 insolence$0 insult$1 lip$2 meanness$2 misbehaviour$0 misconduct$1 mouth$3 pertness$0 ribaldry$0 sauce$0 sauciness$0 vulgarity$0

35 MAKE KNOWN(v): advertise$0 advise$2 air$7 announce$1 blazon$0 circulate$1 communicate$0 convey$2 declare$2 disclose$1 divulge$0 expose$2 express$2 impart$1 intimate$6 introduce$1 leak$5 mention$1 post$2 proclaim$0 promulgate$0 propagate$2 publicize$0 push$4

release$3 reveal$1 say$2 show$1 speak$1
spread$3 tell$1 uncover$2 unfold$2 unveil$0
ventilate$0

41 REASONABLE(adj): common-sensical$0 considerable$1
credible$1 decent$2 economic$4 economical$3
enlightened$0 equitable$0 fair$3 feasible$0
inexpensive$0 judicious$0 just$3 justifiable$0
legitimate$2 level-headed$0 likely$3 logical$1
logical$2 low$10 lucid$4 moderate$1 normal$2
open-minded$0 plausible$0 presumptive$2
probable$0 rational$1 respectable$2 restrained$1
right$4 sane$2 sensible$1 sober$2 sound$8
temperate$2 tenable$0 thinkable$0 warrantable$0
well-balanced$2 wise$1

According to the authors of CT, there are two criteria by which a word is chosen as an entry: The first is "if it is likely to be looked up as an entry in its own right." Thus, the authors explain, rare or obsolete words do not appear as entries although they may be given as synonyms for other, simpler words. The second criterion is that concrete words are usually not selected as entries, unless they have "genuine synonyms or give rise to a figurative use." The first criterion could explain why phrases, such as *prime mover* or *make known*, do not occur as entries. The second probably applies to *ridge*, which has only concrete senses. However, many words in the sample do not appear to fit these criteria. Lexicographers may choose to review all terminal nodes at once; only those whose number of occurrences exceeds a certain threshold; non-phrasal terminals or non-inflected terminals.[8]

When examining terminal nodes, lexicographers may want to consult another list in parallel: that of all the words which are entries but which do not occur as synonyms. There are some 900 CT-entries that never occur as synonyms. Here are ten:

ABAFT(adv): aft$99 astern$99 behind$

ABDUCTION(n): carrying_off$99 kidnapping$99 seizure$1

ABSENTLY(adv): absent-mindedly$99 abstractedly$99 bemusedly$99 blankly$99 distractedly$99 dreamily$99 emptily$99 heedlessly$99 inattentively$99 obliviously$99 unconsciously$99 unheedingly$99 vacantly$99 vaguely$0

AND(conj): along_with$99 also$ as_well_as$ furthermore$ in_addition_to$ including$ moreover$ plus$ together_with$99

ATHLETE(n): competitor$0 contender$99 contestant$0 games_player$99 gymnast$99 player$1 runner$1 sportsman$99 sportswoman$99

AWE-STRICKEN(adj): afraid$1 amazed$99 astonished$99 awed$99 awe-inspired$99 cowed$99 daunted$0 dumbfounded$0 fearful$1 frightened$0 horrified$99 impressed$99 intimidated$99 shocked$99 struck_dumb$99 stunned$0 terrified$0 wonder-stricken$99 wonder-struck$99

BEDCLOTHES(n): bedding$99 bed_linen$99 blankets$99 coverlets$99 covers$99 sheets$99

FAIR-AND-SQUARE(adj): above_board$99 correct$3 honest$2 just$1.2 on_the_level$ straight$4
FEATURED(adj): given_prominence$99 headlined$99 highlighted$99 in_the_public_eye$99 presented$99 promoted$99 recommended$99 specially_presented$99 starred$99
FEATURING(adj): calling_attention_to$99 displaying$99 drawing_attention_to$99 giving_a_star_role$99 giving_prominence_to$99 giving_the_full_works$99 highlighting$99 making_the_main_attraction$99 presenting$99 promoting$99 pushing$ recommending$99 showing$ showing_off$99 starring$99 turning_the_spotlight_on$99

Turning some terminals into entries may result in a need to turn some of these 900 entries into terminals to maintain consistency.

Vocabulary Inconsistencies

A small percent of the terminal nodes in CT is due to vocabulary inconsistencies. For example, *record* has *annals, archives* and *diary* as synonyms; whereas *annals* and *archives* have the plural *records*; and *diary* has the phrase *daily record.* This inconsistency results in both *records* and *daily record* becoming terminal nodes whereas, it would seem that they should not be, since they are equivalent to the main entry *record.* Identifying this category of terminals is particularly important because its correction involves changes in several entries.

The first category of vocabulary inconsistencies variation is number, that is, cases when the same word-sense is referred to sometimes in the plural and sometimes in the singular. We identified these automatically by running our UDICT morphological analyzer (Byrd 1986) on all the terminals found in noun entries, and retrieving all terminals that are plural forms of English nouns. Here is a sample of twenty nouns:

SAFEGUARDS: security$2
SALES: commercial$1
SALTS: laxative$0
SALUTATIONS: greeting$2 greetings$0 regard$10 regards$0 respect$4 respects$0
SANDS: beach$0 shore$1
SAWBONES: physician$0
SAWS: lore$1
SAYINGS: lore$1
SCHOOLDAYS: childhood$0
SCIONS: issue$7 posterity$1 progeny$0 seed$3
SCORES: a_lot$0 lot$4 lots$0 many$2 multiplicity$0 myriad$2
SCOURINGS: dregs$1 garbage$2 swill$3
SCRUPLES: conscience$1 hesitation$2 morals$0 principle$3
SEATS: seating$0
SECURITIES: holdings$0
SENSITIVIES: feelings$0
SERVANTS: retinue$0

SERVICES: liturgy$0 military$2
SHADOWS: darkness$1 obscurity$2 shade$1
SHEETS: bedclothes$0

The second step was to check if the singular forms of these plural nouns were CT-entries, and if so, whether their synonym lists included any of the entries which listed the plural forms. From the sample, the following five entries were found:

SAFEGUARD$2: ... security$2 ...
SCORE$3: ... lots$0 ...
SCRUPLE$2: ... hesitation$0 ...
SERVICE$4: ... liturgy$0 ...
SHADOW$1: ... darkness$1 ... obscurity$2 ... shade$1

10% of the plural terminal nodes in CT are similar to the five above, in that they have corresponding singular entries whose synonyms intersect with the entries in which the terminals occur. The lexicographer may want to distinguish these cases, where the singular and plural are synonyms (at least, on one sense), from words such as *salts* and *sawbones*, where the singular and the plural differ in meaning. For the synonymy case, a uniform marking convention for both headwords and synonym tokens will be useful. If the senses in question are written as *safeguard(s)* or *shadow(s)*, for example, the synonymy of the two forms is always apparent.

Another type of vocabulary inconsistency is the variation between a single word and a phrase containing the word and a modifier, as in *daily record*. Here we checked all the CT-terminals for two-word phrases composed of a modifier and a head (capitalized) as follows: ADJECTIVE_adverb, adverb_ADJECTIVE, adverb ADVERB, adjective_NOUN, VERB_prep and VERB_adverb.[9] The following is a sample of 20 combinations retrieved in this search:

SCARED_stiff: frightened$0 panic-stricken$0 petrified$2 terrified$0
slightly_DRUNK: tipsy$0
slightly_WARM: tepid$1
unbearably_HOT: scorching$0
unduly_QUICK: hasty$3
vastly_SUPERIOR: overwhelming$0
scarcely_EVER: rarely$1 seldom$0 uncommonly$1
very_MUCH: awfully$2 by far$0 by half$0 considerably$0 dearly$1 far$2 far$3 greatly$0 half$4 heavily$7 highly$1 mightily$1 overly$0 well$8
very_NEARLY: practically$1
very_OFTEN: frequently$0
very_WELL: intimately$1 intimately$2 swimmingly$0
sanitary_MEASURES: hygiene$0
scenic_VIEW: panorama$1
secret_MEETING: assignation$1
secret_PLACE: hide-out$0
semiprecious_STONE: gem$1
servile_FLATTERY: adulation$0
sexual_ACT: intercourse$2
SAW_down: cut$3
SCARE_off: intimidate$0

The second step was to check whether the head in isolation was a CT-entry, and if so, whether its synonyms included any of the entries linked to its corresponding phrase:

SCARED$0(adj): frightened$0 ... panic-stricken$0 ... petrified$2 ... terrified$0
DRUNK$1(adj): ... tipsy$0 ...
WARM$1(adj): ... tepid$1.2 ...
HOT$1(adj): ... scorching$0...
QUICK$1(adj): ... hasty$1 ...
MUCH$2(adv): ... considerably$0 ...
NEARLY$0(adv): ... practically$1 ...
OFTEN$0(adv): ... frequently$0 ...
WELL$5(adv): ... intimately$1.2 ...
VIEW$1(n): ... panorama$1 ...
MEETING$1(n): assignation$1 ...
FLATTERY$0(n): adulation$0 ...
SCARE$1(v): ... intimidate$0 ...

Here, too, we suggest a marking convention for making these entries more consistent. Phrases that are synonymous with their heads can be written as *(slightly)_drunk* or *(unbearably)_hot*. Written in this way, the cross-reference to the single-word entry or synonym remains transparent. The use of parentheses can help to differentiate these phrases from others, such as *secret_places* or *sexual_act*, which are not synonymous with their heads.

Many verbal or adjectival phrases (VERB_prep, PARTICIPLE_prep or ADJECTIVE_prep) in CT occur in run-on entries that themselves consist of the single main-entry word followed by a preposition. Most (but not all) synonyms offered for such run-on entries are phrases, whereas most (but not all) synonyms offered for single main entries are single words. The following entries illustrate this contrast:

UNFAMILIAR:
> >1 alien$1 curious$3 different$4 little_known$99 new$1 novel$1 out-of-the-way$2 strange$2 unaccustomed$2 uncommon$1 unknown$1 unusual$0
> >2 with *with*: a_stranger_to$99 inexperienced_in$99 unaccustomed_to$0 unacquainted$99 unconversant$99 uninformed_about$99 uninitiated_in$99 unpractised_in$99 unskilled_at$99 unversed_in$99

UNACCUSTOMED:
> >1 with *to*: a_newcomer_to$99 a_novice_at$99 green$3 inexperienced$0 not_given_to$99 not_used_to$99 unfamiliar_with$0 unpractised$99 unused_to$0 unversed_in$99
> >2 new$1 out_of_the_ordinary$0 remarkable$0 special$1 strange$1.2 surprising$0 uncommon$1 unexpected$0 unfamiliar$1 unprecedented$0 unusual$0 unwonted$0

INEXPERIENCED:
› › 0 amateur$ callow$0 fresh$7 green$3 immature$1 new$1 raw$4 unaccustomed$1.2 unacquainted$99 unfamiliar$1 unfledged$0 unpractised$99 unschooled$99 unseasoned$99 unskilled$0 untrained$0 untried$0 unused$1 unversed$99 wet_behind_the_ears$0

In our processing of CT, we have duplicated run-on entries, so that our version of CT has *unfamiliar$2* also referenced as *unfamiliar_with$0* and *unaccustomed$1* also as *unaccustomed_to$0*.

Let us now examine the synonym lists for these entries. A distinction between the sense of the single adjective and the sense of the adjectival phrase is made, as can be seen from the separate links between *unfamiliar_with* and *unaccustomed_to* on one hand and *unfamiliar$1* and *unaccustomed$2* on the other. However, the distinction appears inconsistent: the simple *unacquainted* and *unconversant* are given as synonyms for *unfamiliar with*. (Why not *unacquainted_with* and *unconversant_in?)* Similarly, the phrasal *inexperienced_in* is given as a synonym of *unfamiliar_with*, but the simple *inexperienced* is given as a synonym of *unaccustomed_to*. There are many other such cases, which could be aided by the lexicographer reviewing a list of all corresponding phrasal and single words.

Other Asymmetries

Of the non-terminal asymmetries, about 18% are instances of hypernymy (the superordinate relation) or hyponomy (the subordinate relation).[10] For example, *book* lists *manual* as a synonym, but *manual* does not list *book*; instead, special types of books, such as *handbook*, are given. This is because *book* is really a hypernym (not a synonym) of *manual*. Hypernym links are truly asymmetric in nature. The lexicographer's view of synonymy will determine whether such hypernym links should be included in the thesaurus, and if so, whether they should be separated from or marked differently than genuine synonyms.

In CT, hypernym links are not distinguished from other links, and so there is no automatic way to retrieve them. The best we could do was to produce an approximate list of hypernym links by comparing CT-synonyms with hypernym and hyponym lists that we have on-line, in our taxonym files. Our taxonym files were built automatically with information extracted from W7. For a given word *a*, the files contain all the words defining it (that is, the words occurring as heads of its definitions) and all the words which *a* defines (that is, words in whose definitions *a* is the head) (Chodorow et al. 1985). Following are some sample results of intersecting the CT synonym lists with our hyponym lists. For each entry on the left, we list its CT-synonyms that were found to be hyponyms of it. Since this is the result of a comparison between two different sources, each with its own sense separation, no sense disambiguation was possible.

TABLE(n):	bench board
TACK(n):	thumbtack
TACKLE(n):	rig
TAINT(n):	spot stain
TALE(n):	romance yarn

TALENT(n):	ability gift
TALK(v):	blather chat gab gossip harangue jaw palaver
TANGLE(n):	snarl
TAP(n):	touch
TART(n):	tartlet
TASK(n):	business chore duty job mission work
TASTE(n):	decorum palate partiality smack

The following is the result of intersecting CT-synonym lists with our *hypernym* lists:

TABLEAU(n):	representation
TABOO(n):	prohibition
TACK(n):	course direction method nail
TACT(n)	perception
TACTIC(n):	device method
TAG(n):	marker
TAIL(n):	end line
TALE(n):	narrative relation report
TALENT(n):	aptitude endowment power
TALK(n):	discussion negotiation
TANGLE(n):	mass
TART(n):	pie

The majority of the non-terminal asymmetries are not instances of hypernymy. For example, *assembly* has *throng* listed as a synonym of one of its senses, but *throng* does not list *assembly* as a synonym, although it does give *assemblage, congregation, multitude*, and other related words. Perhaps many of these omissions are due to the fact that rare, very formal or metaphoric words tend not to be offered as synonyms. This may explain why *conversant, familiar* and *informed*, for example, are listed as synonyms of *cognizant*, while *cognizant* is not listed as their synonym. Another possible reason could be cases when a central sense of one word is synonymous with a very peripheral sense of another. One sense of *say* lists *add*, as in "He added that he would do the demonstration." The entry for *add* does not, however, contain this peripheral sense and deals only with the arithmetic sense of *add* and the sense of enlargement. Unfortunately, it is not evident how to automatically produce a list of asymmetries due to these reasons.

Intransitive Links

In the discussion of asymmetry in CT we presented various lists of word-senses which we propose as candidates for addition or deletion. In this section, we briefly present a tool that may assist lexicographers in reclassification. In our manipulation of the synonymy links in CT we have been building synonym trees, with a process called SPROUTING. A sense of a headword is chosen as the root of the tree (for example, *house1*); a program called SPROUT (Chodorow et al. 1985) starts with the root node and retrieves from the thesaurus all of its synonyms. These word senses are the first-level descendents (daughters) of the root. SPROUT then applies recursively to each of the daughter nodes, generating their daughters, etc. In this

way, the tree is generated in a breadth-first fashion. The process is complete when the only nodes that remain open are either terminals (i.e. nodes that have no daughters) or nodes that appear earlier in the tree, indicating a cyclic structure. Because of the structure of CT, trees of this kind reach closure only after picking up most of the CT tokens. The *house1* tree, for example, contains 85% of the total number of noun senses.

In an attempt to maintain semantic content, we have explored ways of automatically pruning the sprout tree when a semantically irrelevant branch is generated. Before any synonym is accepted as a node of the tree, its descendents are checked against the immediate descendents of the root node. If the intersection of their mutual synonym lists is not null, the node is accepted into the sprout tree. For lexicographers, the nodes that are rejected can be helpful in detecting faulty links. Of particular importance are the nodes that point back to different senses of nodes already encountered. For example, the following branch of the *house1* tree points to a problem:

house1 — > building1 — > construction1 — > building2

We have noticed that in most such loops, the problem lies in poor sense separation in the original CT entries. *Building2*, for example, is a mixture of the act of building, the object built and its design. We do not recommend an exhaustive review of all such loops. The task seems too formidable—we found 260 loops in the first 1000 entries—but the availability of the sprouting mechanism may be useful when extensive changes are entertained for a family of word senses.

Conclusion

In this paper we have discussed various types of asymmetric and intransitive links found in THE NEW COLLINS THESAURUS (CT). We believe that the existence of these links is typical of most thesauri. Since it is the result of a vast number of individual decisions taken by one or several lexicographers in often conflicting situations, some degree of inconsistency is inevitable. We have shown how our computer programs can provide lexicographers with various sorted listings of these links, so that the process of reviewing and correcting the inconsistencies can be significantly facilitated.

Notes

[1] We have stored CT as a DAM file (Byrd et al. 1986) with 16,794 entries containing a total of 287,136 synonym tokens.

[2] Part-of-speech information was obtained with the UDICT computerized lexicon system (Byrd 1986).

[3] A notable exception is ROGET'S II, NEW THESAURUS, in which all members of a synonym set are symmetrically and transitively linked.

[4] It should be noted that CT's vocabulary is limited. Thus, it does not contain the verb "perk" or the noun "saw" as an instrument of cutting. The list of transitive and symmetric sets will vary with the size of the on-line source.

[5] The percentage of non-transitive links does not include synonyms which have no entries in CT (see the section on terminal nodes); nor does it include synonyms which could not be

disambiguated (see the section on sense disambiguation). Thus 65% is a conservative estimate.

[6] The number following the dollar sign indicates the sense number. No number indicates that the intersection is null and therefore a sense number was not picked up. 99 indicates that the word has no entry in CT and consequently no sense numbers. 0 means that there was only one sense given in the entry.

[7] The total of 221,957 represents the number of non-terminal links, as discussed in the following section on terminal nodes.

[8] It is interesting to note that the infrequently occurring terminals do not differ markedly from the frequently occurring ones.

[9] We assumed that combinations of the form noun NOUN_have meanings that are distinct from the meaning of their heads in isolation.

[10] The following percentages were computed on the basis of fifty random entries.

References

Cited Dictionaries

THE NEW COLLINS THESAURUS (CT). 1984. Glasgow: Collins Publishers.

ROGET'S II, THE NEW THESAURUS. 1980. Boston, MA: Houghton Mifflin.

WEBSTER'S SEVENTH NEW COLLEGIATE DICTIONARY (W7). 1963. Springfield, MA: Merriam.

Other Literature

Byrd, R. J. 1983. 'Word Formation in Natural Language Processing Systems.' in *Proceedings of IJCAI—VIII*, 704—706.

Byrd, R. J. 1986. 'Dictionary Systems for Office Practice' in *Proceedings of the Grosseto Workshop "On Automating the Lexicon"*. Also available as *IBM Research Report* RC 11872.

Byrd, R. J., G Neumann, and K. S. B. Andersson. 1986 'DAM—A Dictionary Access Method', *IBM Research Report*.

Chodorow, M. S., R. J. Byrd, and G. E. Heidorn. 1985. 'Extracting Semantic Hierarchies from a Large On-Line Dictionary' in *Proceedings of the Association for Computational Linguistics*, 299—304.

Chodorow M. S., Y. Ravin, and H. Sachar. 1988. 'A Tool for Investigating the Synonymy Relation in a Sense-Disambiguated Thesaurus' in *Proceedings of the Second Conference on Applied Natural Language Processing*, 144—151.

Katz, J. and J. Fodor. 1963. 'The Structure of a Semantic Theory' in *Language* 34: 170—210.

Lesk, M. 1986. 'Automatic sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone' in *Proceedings of 1986 SIGDOC Conference*. Canada.

Quine, W. 1960. *Word and Object*. Cambridge, Massachusetts: MIT Association for Computing Machinery.