The Influence of Corpora on Lexicons: Corpora Use in the Creation of COMLEX Syntax and NOMLEX

Catherine MacLEOD and Ralph GRISHMAN, New York, USA

Abstract

It is now generally accepted that a text corpus plays an important role in the production of hard-copy dictionaries. In this paper, we discuss the influence a corpus can have on the creation of lexical resources for computer use. In the creation of COMLEX Syntax and NOMLEX, two on-line lexicons produced by the authors at New York University, we used two different corpora, one composed of a small (one million words) balanced corpus (the Brown Corpus) plus a large amount of newspaper data and the other, a large balanced corpus (100 million words) of British English (the British National Corpus). We point out how the use of these two corpora affected the resulting lexicons in different ways and to differing degrees and we suggest what we feel would have been the ideal corpus for our purposes.

1 Introduction

In the development of our two machine-readable dictionaries (COMLEX Syntax and NOMLEX), we used two corpora, one a hybrid consisting of the Brown Corpus plus a large amount of newspaper text and the other the British National Corpus (BNC). We will discuss the necessity of using corpora and the advantages and disadvantages of these two very different corpora. Lastly, we will discuss the type of corpus that we consider would have made a considerable improvement in the creation of our lexicons.

2 COMLEX Syntax

COMLEX Syntax [Macleod et al. 1997] is a large (over 39,000 head words) syntactic dictionary developed at New York University (NYU) under the auspices of the Linguistic Data Consortium (LDC) and available from the LDC for both commercial and research use. This dictionary was intended for use in natural language processing (NLP) primarily as part of a system for parsing texts. It assigns to the major parts of speech (noun, adjective, verb and adverb) a rich set of syntactic classes and features, including detailed information on complements for verbs, nouns and adjectives.

3 Use of Corpora in COMLEX Syntax

The corpus used while creating COMLEX contains about 100 megabytes of text including most of the Brown Corpus (7 MB), Wall Street Journal (27 MB), San Jose Mercury (30 MB), Associated Press (29.5 MB) and miscellaneous selections from the Treebank Literature (1.5 MB). The sections omitted from Brown were portions which contained non-standard English, including poetry and pieces of “Tom Sawyer”, which makes rich use of the vernacular. We had a number

of linguistics graduate students entering the COMLEX word classes, using an entry program developed at NYU. This program includes a concordance taken from our corpus, which displays citations of the word being entered. The elf (Enterer of Lexical Features) was able to look at as many examples as necessary to augment his/her understanding of the syntactic patterns that co-occurred with the particular word. The elves had access to hard copy dictionaries and their own expertise as native speakers of English, but the concordance served to give actual examples of intuitive choices and as reminders of cases which the elf might have ignored. The use of corpora in dictionary building has become generally accepted as being the best way to capture actual usage rather than the knowledge/experience of the individual lexicographer. The use of an on-line concordance was essential in aiding the coverage, as well as the accuracy, of COMLEX.

3.1 Use of a Corpus in Tagging

As part of COMLEX, we have provided 100 corpus citations of each of 750 'high frequency verbs'. Each citation was tagged with its COMLEX complement class. This tagging provides useful information on the distribution of complements for both sentence analysis and generation. The high-frequency verbs were selected based on the part-of-speech tagged subset of our corpus (the "POS corpus")[1].

It was in the tagging phase of COMLEX that the inadequacy of our corpus was clearly demonstrated. First of all, the make-up of the POS corpus, with its preponderance of newspaper text, skewed the choice of high-frequency verbs. This can be seen by comparing the frequency-ranked list from this corpus with that from Brown, a more balanced corpus. Among the top 50 verbs from our corpus, quite a few (business-related) verbs were not in the top 50 from Brown, including *sell, rise, buy, pay*, and *increase*. In fact, some were not even in the top 750 from Brown, such as *post, boost, invest, value*, and *resign*.

The other shortcoming of our corpus was seen during the actual tagging. We tried to lessen the effect of our unbalanced corpus by choosing citations from the Brown corpus first. Only if there were not enough examples (100) for that verb would we start tagging in the Wall Street Journal. We obtained slightly more than half of our citations from Brown, about one quarter from the Wall Street Journal and the other quarter mostly from the San Jose Mercury with 3% from the Associated Press and a negligible amount from miscellaneous treebank literature.

During our tagging, we ran across a number of complements which were not part of the COMLEX inventory of complements and seemed in fact not to be common in "general" English. Among these were a variety of complements containing NUNITP (NP constructions containing units) as demonstrated in Table 1[2].

COMLEX Complement	Example Sentence
NUNITP	- buys his suits four at a time at Neiman-Marcus in Dallas and PAYS as much as $250 each.
FROM-RANGE	- occupancy rates in major hotels here RANGED from 48 to 74 percent last year.
NP-NUNITP-TO-RANGE	The ordinance would INCREASE fees from $1 for males and $2 for females to a flat $5 a dog.
NUNITP-TO-RANGE	The payroll tax would actually RISE to 7.5 per cent starting Jan. 1, 1963.

Table 1: Examples of NUNITP complements

We did a small study to measure the degree to which text type affected the distribution of these complements [Macleod et al. 1994]. Of the 43 verbs which had NUNITP complements, 21 did not appear at all in Brown. Another 8 verbs were very high frequency verbs where all 100 citations were taken from Brown. The frequency of NUNITP complements on these verbs ranged from 1% to 8%. The most interesting results of this study were found in the verbs where a substantial number of citations came from both Brown and the Wall Street Journal (wsj). The distribution of these complements in the two texts was seen to be quite different. See Table 2.

verb	source	complement frequency	
jump	57 from brown	2 nunitp-to-range	4%
	43 from wsj1	21 nunitp-to-range	49%
advance	42 from brown	2 nunitp-to-range	5%
	58 from wsj1	36 nunitp-to-range	
		1 np-nunitp-to-range	64%
climb	63 from brown	0 nunitp-to-range	
		2 nunitp-pp	3%
	37 from wsj1	26 nunitp-to-range	
		0 nunitp-pp	70%
range	63 from brown	16 nunitp-from-range	25%
	37 from wsj1	20 nunitp-from-range	54%
quote	44 from brown	0 np-at-nunitp-pred	
	56 from wsj1	35 np-at-nunitp-pred	62%
	97 from BNC	0 np-at-nunitp-pred	

Table 2: NUNITP Complement Distribution in Brown and the Wall Street Journal (Percentages represent the fraction of the NUNITP complements which occur in each corpus)

In order to ascertain that this different distribution was real, we also looked at examples for one verb, *quote*, from the British National Corpus (BNC)[3]. The distribution of complements is shown in Table 3[4]. As can be seen, Brown and the BNC are remarkably consistent as to complement distribution, whereas the Wall Street Journal has an NUNITP complement for over half of the instances of *quote*.

Complement	BNC (97)	brown (44)	wsj1 (56)
NP-AT-NUNITP-PRED	0	0	35
NP	47	20	10
NP-PP	6	1	1
NP-AS-ING	9	11	8
NP-AS-ADJP	1	0	0
NP-AS-NP	0	1	0
NP-TO-NP	3	0	0
PP (from)	8	4	0
VSAY	11	0	0
NP-VSAY	4	3	0
PP-VSAY	0	3	0
PARENTHETICAL	6	1	0
NP-THATS	2	0	0
NP-ING-OC	0	0	1

Table 3: Tagged COMLEX-Syntax complements

3.2 Use of the British National Corpus in classifying Adverbs

The classification of COMLEX adverbs is different from that of the other parts of speech. We do not assign complements to adverbs, but rather classify them positionally. The use of a corpus is essential in identifying the possible positions of an adverb. This turned out to be particularly problematic for infrequent adverbs, some of which did not occur at all in our corpus. Therefore, we used the BNC for our reference. In Table 4, we demonstrate the importance of having a balanced corpus for adverb classification. Adverbs which are not unusual often do not occur in our corpus (even though it is a large corpus); however, they can be found in the BNC. Even rare adverbs will have some examples to look at. Without access to the BNC, we could not have entered these adverbs at all.

Adverb	COMLEX Corpus	BNC
tactfully	0	162
humorously	0	51
unarguably	0	15
mindlessly	1	26
pejoratively	1	11

Table 4: Adverb Frequency in The COMLEX Corpus and the BNC

4 NOMLEX

NOMLEX [Macleod et al. 1998] is a dictionary of nominalizations. It is based on COMLEX Syntax verb complements and relates the argument structure of its associated verb to the parts

of the nominalization phrase. We used both our corpus and the BNC for examples. Although the differences were not as clear for the nominalizations, we found examples that underlined the usefulness of a large balanced corpus. *Deduction* is the homograph nominalization for both *deduct* and *deduce*. In our corpus *deduction* appeared 278 times; *deduction from* appeared three times and was always *deduct from*. *Deduction that* was always a relative clause on *deduct*. Instead in the BNC (411 instances of *deduction*) more than 1/3 of the instances of *deduction from* come from *deduce* and 3/4 of the instances of *deduction that* are from *deduce*.

The argument structure of *account* differed from our corpus and the BNC, as well. In 50 instances of *account to* from our corpus, only one *to* was an argument of *account* and twenty-nine instances were *award account to*. In 50 random examples of *account to* from the BNC, 4 represented the argument *to* while one only was *award account to*.

5 The Relative Worth of Different Corpora in Dictionary Creation

We have seen above that the COMLEX and NOMLEX projects made heavy use of corpora during their construction, both for entering and tagging. We found the BNC to be preferable when we compared it to the Brown Corpus and the Wall Street Journal; it patterned very like the Brown Corpus as we saw in Table 3. It is clear from the discussions above that the BNC would have been the best possible corpus for dictionary creation. This is borne out also by the fact that The FrameNet project [Baker et al. 1998], a syntactic/semantic network now being built at Berkeley, is using the BNC. An important part of Fillmore's FrameNet project is the tagging of corpus examples with syntactic and semantic frames for which they reference the BNC.

However, there is a disadvantage to this corpus for those of us who deal in American English. Although the balance and scope of this corpus is better than any other corpus available for dictionary work[5], the corpus is of British English.

This may not seem a problem to those who see the difference as being confined to a small number of lexical items but the truth is otherwise and more serious. The grammar of American English (A.E.) varies from British English (B.E.) quite significantly. For example, British English often makes use of a to-infinitive complement where American English does not. In the following examples from the BNC, *assay*, *engage*, *omit* and *endure* appear with a to-infinitive complement; there were no examples found in our corpus of this construction although the verbs themselves did appear. For the first two verbs, one can argue that there is not an equivalent verbal meaning in A.E. but, for the last two, the meaning can be paraphrased in A.E. by the gerund, as seen in Table 5. Note that the B.E. examples are from the BNC and the A.E. examples are paraphrases.

Verb	Eng.	BNC ID	Example sentences
assay	B.E.	G0M 2038	Jerome crept to the foot of the steps, and there halted, baulked, rather, like a startled horse, drew hard breath and ASSAYED TO MOUNT, and then suddenly threw up his arms to cover his face, fell on his knees with a lamentable, choking cry, and bowed himself against the stone of the steps.
engage	B.E.	E9V 768	A magnate would ENGAGE TO SERVE with a specified number of men for a particular time in return for wages which were agreed in advance and paid by the Exchequer.
omit	B.E. A.E.	FS4 941	'What did you OMIT TO TELL your priest?' "What did you OMIT TELLING your priest?"
endure	B.E. A.E.	CD2 1061	But Carteret's wife, who frequented health spas, could not ENDURE TO LIVE with him or he with her: there were no children. But Carteret's wife, who frequented health spas, could not ENDURE LIVING with him or he with her: there were no children.

Table 5: Examples of B.E. verbs followed by to-infinitives

Verb complementation containing prepositions often differs from B.E. to A.E. John Algeo [Algeo 1988] gives a number of examples. In B.E., *cater for* and *cater to* both occur but *cater to* has a pejorative connotation and is less frequent. In A.E., only *cater to* is used and is not considered pejorative. B.E. *claim for* contrasts with A.E. *claim* + NP (claim for benefits vs claim benefits) and conversely *sound* + NP is acceptable in B.E (that sounds a good idea) but not in A.E. which demands the preposition *like* (that sounds like a good idea).

Adverbial usage is also different. The B.E. use of *immediately* in sentence initial position, is not allowed in A.E. For example, B.E. *Immediately I get home, I will attend to that.* is incorrect in A.E. where we would say *As soon as I get home, I will attend to that.* We do concur in the example *I expect him to go immediately.* which is correct in either language.

Other syntactic differences are the formation of questions with the main verb "have". In B.E., one can say, "Have you a pen?" where A.E. speakers must use "do" ("Do you have a pen?"). Support verbs for nominalizations also differ. Note the B.E. "take a decision" vs the A.E. "make a decision".

With these considerable differences and the fact that lexical items may be over- or under-represented or not present at all, it is clear that what is needed is a large balanced corpus of American English on the lines of the BNC but from American texts. The last effort to make a balanced generally-available corpus of American English was the Brown Corpus. This is an excellent corpus but it is too small (one million words compared with 100 million in the BNC) and somewhat out of date (having been constructed in the 1960's).

6 An American National Corpus

In this paper, we have discussed our experience using different corpora for creating and tagging COMLEX Syntax and NOMLEX. An unbalanced corpus skews the data towards whichever type of text predominates, making a general resource very hard to construct. The Brown Corpus is balanced but is too small for many lexicographic purposes and (a more minor concern) it is over 30 years old. The BNC is large and balanced but, unfortunately for those working in American English, deals with British English texts. What we are now awaiting is the creation of an American National Corpus (ANC) with the size and balance needed for American English lexicography.

At the first Language Resources and Evaluation Conference in 1998, a proposal was made for such a corpus, containing at least 100 million words of American English, balanced much on the lines of the BNC [Fillmore et al. 1998]. A committee of researchers and lexicographers[6] is now endeavoring to make this corpus a reality, freely available to all researchers in the near future. We are now gathering a Consortium to build the ANC. This consortium will provide minimal funding, texts and advice for the creation of the base corpus. The automatic annotation and the distribution of the ANC will be handled by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. The base corpus will be useful for lexicographers who need examples of usage over a broad area. NLP researchers, especially those involved in statistical studies, will need a more carefully annotated corpus. This will be done in a second stage, which because of the great cost of hand annotation will have to be supported by government funding agencies. However, by using the ANC we will be able to eliminate the problems of *taking* rather than *making* our decisions.

7 Summary

In our experience creating dictionaries for the use of natural language processing, we find that the use of a corpus is a necessity. Unfortunately, the available corpora are inadequate either because they are small, are not balanced, are not available to the general research community (for example the many “in-house” corpora created by publishing companies for their exclusive use) or do not contain texts in American English. In fact, there is no present corpus which meets our needs; that is why we and others are involved in creating the resource that we (and, we believe, many others) need.

8 Acknowledgements

The work on COMLEX-Syntax was supported by the Advanced Research Projects Agency through the Office of Naval Research under Awards No. MDA972-92-J-1016 and N00014-90-J-1851, and The Trustees of the University of Pennsylvania. NOMLEX was created under the National Science Foundation Grant No. IRI-9633286. The initial organizational meeting for the American National Corpus held at Berkley was funded by the National Science Foundation grant ISI-9978422.

Notes

[1]This corpus consisted of four Penn Treebank files: miscellaneous (without Dubois poetry and Tom Sawyer), Brown, Department of Energy documents, and the Wall Street Journal.

[2]These are actual tagged examples from our corpus.

[3]Thanks to Lou Burnard of Oxford University, we were able to access the BNC via the Web.

[4]Because of limitations on our access, we could not simply obtain 100 consecutive citations from the BNC of any form of the verb *quote*. We obtained instead 50 random instances of *quote* as a base form verb (3 were unusable), and 50 instances of *quoted* as a past tense verb.

[5]This obviously does not include "in-house" corpora which are generally not available outside the company or institution that developed them.

[6]The committee includes Frank Abate, Charles Fillmore, Ralph Grishman, Nancy Ide, Daniel Jurafsky, Mark Liberman, Catherine Macleod and Wendalyn Nichols.

References

Algeo, John. British and American Grammatical Differences, in: *International Journal of Lexicography*, Vol. 1 No. 1, 1988, pp. 1-31.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project, in: *The Proceedings of COLING-ACL '98,* Montreal, Canada, August, 1998, pp. 86-90.

Fillmore, Charles, Nancy Ide, Dan Jurafsky and Catherine Macleod. An American National Corpus: A Proposal, in: *The Proceedings of the First International Conference on Language Resources and Evaluation* Granada, Spain May, 1998, Poster Session 3: Corpus pp. 965-970.

Macleod, Catherine, Ralph Grishman, and Adam Meyers. Developing Multiply Tagged Corpora for Lexical Research, in: *The Proceedings of the Post-Coling International Workshop on Directions of Lexical Research*, Beijing, China, August, 1994, pp. 11-22.

Macleod, Catherine, Ralph Grishman and Adam Meyers. COMLEX Syntax, in: *Computers and the Humanities*, Vol. 31 No. 6, 1997/1998 pp. 459-481.

Macleod, Catherine, Ralph Grishman, Adam Meyers, Leslie Barrett, Ruth Reeves. NOMLEX: A Lexicon of Nominalizations, in: *Proceedings of EURALEX'98*, Liege, Belgium, August 1998, pp. 187-193.