

"Corpus linguistics" or "Computer-aided armchair linguistics"

Charles J. Fillmore

Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations.)

Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations.)

These two don't speak to each other very often, but when they do, the corpus linguist says to the armchair linguist, "Why should I think that what you tell me is true?", and the armchair linguist says to the corpus linguist, "Why should I think that what you tell me is interesting?"

This paper is a report of an armchair linguist who refuses to give up his old ways but who finds profit in being a consumer of some of the resources that corpus linguists have created.

I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body.

During the early decades of my career as a linguist, I thought of myself as fortunate for having escaped corpus linguistics. Of course, I wouldn't have used the term corpus linguistics in describing my good fortune: maybe I would have called it statistical linguistics.

The situation was this. When I showed up as a beginning graduate student at the University of Michigan's linguistics program, a long time ago, the first person I considered as a possible dissertation director was the kind of professor I myself would like to be able to be, namely, someone with a well-articulated research agenda who asked each of the students who came under his wing to take on a predetermined assignment within that agenda. If I wanted him to be my mentor, I was to carry out the following assignment.

First, I was to make extensive tape recordings - actually, at the time, it may have been wire recordings - of natural conversations in English and Japanese. After doing that, I was to

choose and justify a set of empirical criteria for phonemic analysis that could be applied to each of these languages.

(Those were the days when, realizing that a single language could be given more than one phonemic analysis, people worried - correctly - that phonemic descriptions of different languages couldn't be considered comparable unless one applied, equally to each of the languages being compared, precisely the same set of decision-making criteria.)

Armed with a carefully justified phonemic analysis, for each language, I was then to prepare phonemic transcriptions of all of the conversations that I had recorded.

That was the first part - maybe a year, maybe a year and a half. The next and more important part of the job was to take from each transcript cumulatively larger samples - say, the first 200 phoneme tokens, the first 400 phoneme tokens, the first 600 phoneme tokens, etc., and with each of these growing samples, to plot out the relative frequencies of the phonemes. I was to continue doing this until I had determined, for each of these languages, the mean length of discourse samples, in terms of stretches of phoneme tokens, at which the relative frequencies of the phonemes stabilized.

If the results, using this measure, turned out to be significantly different for English and Japanese, and if I could argue that such a difference could be related to, say, phonotactic characteristics of the two languages, then the results of the research could be seen as contributing, to phonological scholarship, some practical guidelines on how large a corpus of spoken language needs to be for it to be considered an adequate reservoir of the phonological phenomena of the language.

I rejected the assignment. But now, having recalled it for the sake of the opening paragraphs of this talk, I find that it doesn't sound quite as bad to me today as it did thirty-some years ago. There have been times when I've regretted the missed opportunity, since I now know that in the process of moving carefully through a text, of any sort, I would undoubtedly have learned a great deal about both of these languages. I must admit, of course, that I can imagine languages for which the relative frequencies of their phonemes would stabilize long before all of their interesting phonological properties had checked in. The fact is, I couldn't really imagine myself becoming interested in such a project; nor could I imagine what I would be able to say in the section of the dissertation that was supposed to bear the title, "The Significance of the Present Research".

The year was 1957. I soon came to be subjected to other intellectual currents within linguistics; and, in fact, before long I was, without the encouragement of my Michigan teachers, converted to a way of doing linguistics which not only did not depend on the careful examination of corpora but whose practitioners often actively ridiculed such efforts.

There were two sorts of activities in those days that would have fit the category corpus linguistics: the first was the study of corpora that field linguists had gathered for poorly documented languages, with both descriptive-linguistic and ethnographic interests in mind; and the second was the study of the statistical properties of languages for which there was no scarcity of data. I was a good disciple, and I learned the correct things to say to linguists" who pushed either of these kinds of studies on me. To the first I learned to say that the knowledge linguists need, in order to come up with an account of a language that met the requirements of a generative grammar, could not be derived from a corpus, however large. For

that we need to appeal to the kind of intuitive knowledge of their language possessed only by native speakers, "the people who know not only what one can say in the language, but also what one cannot say. And as long as we've got that, we don't need anything else.

To the second group of linguists I learned to quote the philosopher Michael Polanyi, author of *Personal Knowledge* (1958), who had said that if natural scientists felt it necessary to portion out their time and attention to phenomena on the basis of their abundance and distribution in the universe, almost all of the scientific community, would have to devote itself exclusively to the study of interstellar dust. And I admired, and later shamelessly imitated, Morris Halle's performance in a debate with policy makers in foreign language, education who sought funding for corpus building so that it could become possible to design programs in which one could teach a language words and structures in the order of their frequency of occurrence in natural texts. Halle said that if driver education were handled according to such principles, nobody would be taught how to put an automobile into reverse gear, since the distance an automobile covers while moving backwards is a hardly-noticeable fraction of the distance it covers when moving forward.

Later on I sometimes found myself arguing with people who were defending the superiority of corpus studies against those who kept pointing out that there were many important features of English that simply were not to be found in the corpora that were then available. I would hear my opponents say that this is a pointless objection: all it means is that we need a larger corpus. But the answer to that was easy: that the ability to judge that some corpus is not large enough to be representative of the phenomena of the language, is an ability based on the recognition that certain things which the linguist, as a native speaker, intuitively knows about the language are not exhibited in the corpus. In the end, there is simply no way to avoid reliance on intuitive knowledge.

The most convincing part of the case for using a corpus was that it makes it possible for linguists to get the facts right. Authenticity was the key word. There was a lot of evidence that linguistic intuition, so-called, isn't always reliable, but what one 'finds in a corpus more or less has to be taken as authentic.

On the question of the authenticity of one's data, I have in recent years been given reason to believe that my own position in linguistics is a confused "one. A few years ago, my (I think) friend William Labov went around the world giving a lecture in which something that I had written was offered as a paradigm example of what he called "woolly minded introspectionism". In attempting to demonstrate certain kinds of fit between linguistic form and aspects of language use, I had suggested that a particular utterance form could not be used over the telephone. My example involved the colloquial gesture-requiring demonstrative *yea*, as in *It was about yea big*. For this sentence, I the addressee has to be watching the speaker (Fillmore 1972). Labov, master observer of language as he is, soon after reading my claim, heard somebody¹ use just that expression over the telephone. I am convinced that the person Labov heard would have corrected himself instantly if he had realized what he had just said, but nevertheless I stand accused and convicted as a woolly minded introspectionist. In a recent meeting with some Soviet linguists I was informed that my work was admired in their group because I always concerned myself with real language as opposed to made-up language; but shortly before that, at a conference of non-generative linguists, after I had

presented the results of some corpus-based research I'd been doing with Japanese, two different members of the audience spoke to me saying almost the same thing, something about how eye-opening it must have been for somebody like me to look at real data!

My own interest in corpora has so far been exclusively in respect to their ability to supply information about lexical or structural features of a language which the usual kinds of accidental sampling and armchair introspection could easily allow us to miss. The kind of work I have in mind proceeds like this. We extract, from a large corpus, passages exhibiting particular phenomena. We do manual processing of these examples: we record observations about them in some sort of structural database; we sort the examples by various criteria, we stare at the groups of examples we have collected, we speculate on relations among the phenomena that we observe, we consult the database in respect to our speculations, and so on. The basic rule is that we make ourselves responsible for saying something about each example that we find.

This is similar in a number of ways to traditional lexicographic methods, working off of a collection of citation slips accumulated by the lexicographer or by members of the dictionary's reading program team. The difference is that - before COBUILD¹ at least - the citation slips the lexicographers examined were largely limited to examples that somebody happened to notice; the corpus work I am talking about here requires a principle of total accountability.

I have worked with on-line corpora on several projects, all of them fairly recently. One involves English conditional sentences, in which I am using mainly brochures from the U.S. Department of Agriculture;² another involves Japanese clause connectives, for which the corpus is a series of textbooks on science used in Japanese middle schools.³ But today I want to discuss two research efforts aimed at the lexical description of two English words, *risk* and *home*.

The *risk* work⁴ which was carried out in collaboration with Beryl T. Atkins, lexicographical adviser at the Oxford University Press, began with a comparison of the *risk* entries - for both the noun and the verb - in ten monolingual English dictionaries, both British and American, and noticing certain discrepancies among them. We decided to find out what a large corpus could show us about the behavior of this word.

In the case of the verb, we can notice that there are three different kinds of direct objects. To see the differences, consider a setting in which we are talking about the advisability of your climbing up a particular cliff. I might tell you that as far as I'm concerned, *I wouldn't risk the climb*. To give a little content to my worries, I warn you that since the cliff is steep and slippery, *You would risk a fall*. To convince you that the matter is serious, I might warn you that *You would be risking your life*. The *climb* names what you might do that could put you in danger. The *fall* is what might happen to you. And *your life* is what you might lose.

The *Collins Cobuild English Language Dictionary* listed all three uses, as did *Longman Dictionary of Contemporary English*, but all the others had only two of them, not always the same two.

Mrs. Atkins had the *risk* KWIC concordances from the Birmingham corpus, but it soon became obvious that to be able to sort the examples according to the senses they exhibited; we needed sentence-long contexts. From IBM Hawthorne we received all of the sentences containing the word *risk* from a corpus they had acquired from the American Publishing

House for the Blind, representing a 25,000,000 word collection of edited written American English. The number of *risk* sentences was 1743.

Since I have been working on a method of semantic description which emphasizes the background conceptual structures for describing word meanings,⁵ the first thing I wanted to do was to characterize situations involving risk.

All situations for which the word *risk* is appropriate are situations in which there is a probability, greater than zero and less than one, that something bad will happen to someone or something. In talking about such a situation we need to be able to identify the individual who is likely to suffer if things go wrong - call that person the Protagonist in a risk scenario - and we need to be able to speak of the bad things that might happen to this individual - let's call that Harm. All risk situations involve the probability that from the point of view of some protagonist something bad will happen.

The Harm could take the form of damage to or loss of something that the Protagonist cares about. We can refer to that as a Valued Possession of the Protagonist, meaning something that the Protagonist cares about which is endangered in the risk scenario.

The probability that something bad will befall a Valued Possession of a Protagonist might, or might not, be the result of some act performed by the Protagonist. We refer to such an act as the Deed. The Protagonist's Deed might be performed in order to achieve some goal. We refer to the goal the Protagonist had in performing the Deed as, simply, the Goal.

We speak of the structure of notions lying behind a linguistic category as making up a "frame", and of its elements as "frame elements". Since some of the frame elements were seen as present in all situations involving risk, and others only in some, we found it necessary to define three slightly different variants, or sub-frames, of the risk frame. The differences among them can be suggested by the following diagrams, adapted from a notation used in mathematical decision theory, in which branches in a directed graph represent alternative futures, and the nodes are either circles, representing chance, or squares, representing choices.⁶

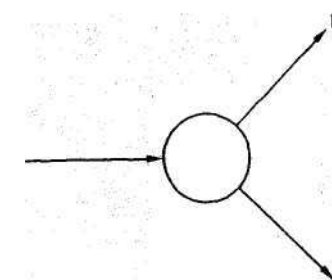


Figure 1.

In Figure 1 we see a situation in which there is the possibility that some harm will occur, but not necessarily as the result of someone's action:

If you stay here you risk getting shot.

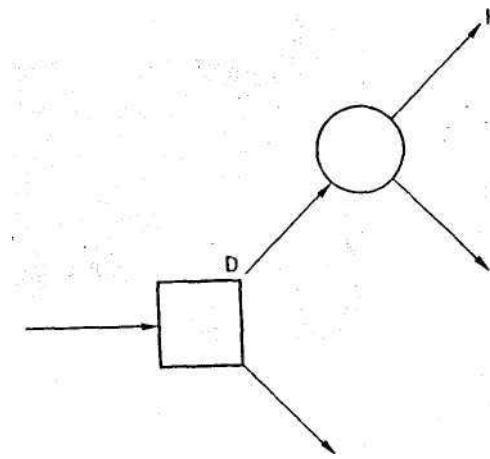


Figure 2.

In Figure 2 we see a situation in which the Protagonist's Deed puts the Protagonist on a path for which there is the possibility of harm:

I had no idea when I stepped into that bar that I was risking my life.

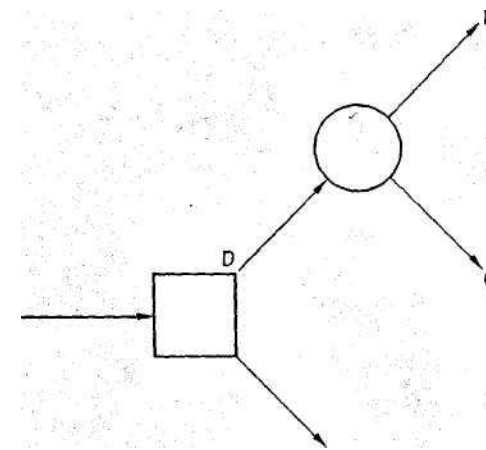


Figure 3.

In Figure 3, the dotted circle - not a standard part of decision theoretic notation - is intended to represent the deliberateness of the Protagonist's decision to perform the Deed. The idea is that the Protagonist chose the path because it is a way of reaching the Goal, while knowing that same path might lead to Harm:

I know I might lose everything, but what the hell, I'm going to risk this week's wages on my favorite horse.

Armed with this set of distinctions we went through all of the verb examples in the corpus and each frame element that got expressed in it. The following is an example of the kind of description this work yields (Figure 4):

When		
you talk like that	Deed	Subclause
you	Protagonist	Subject
risk		
losing your job	Harm	Gerund

Figure 4. "When you talk like that you risk losing your job."

All of the examples of *risk* were transitive. We found NP objects of the verb representing Deed, Harm, or Valued Possession.

Most of us decided to risk *the venture*. <D>

You would risk *death* doing what she did. <H>

Now he was prepared to risk *his good name*. <VP>

In the case of the Harm and Deed frame elements, we also found gerundial objects. In the Deed case the gerund was always a verbal gerund; in the Harm case there were also instances of clausal gerunds.

He risked *committing grave mistakes*. <H>

He had to risk *Pop getting mad at him*. <H>

She risked *going to the pool alone*. <D>

Almost all of the sentences in the corpus could be accounted for, in the sense that we could fit all of their complements and adjuncts into our view of the risk scenario, but there were a few hold-outs, sentences containing syntactic units whose interpretations didn't directly or simply fit into the risk frame. It was the corpus that forced us to deal with these examples, because I am very sure we would not have thought of them on our own. I am referring to adjunct prepositional phrases with *in*, *on* and *to*. Examples:

Roosevelt risked fifty thousand dollars *in* Dakota ranch lands.

You risked a month's earnings *on* that stupid horse!

The captain risked his ship *to* torpedo attack.

Risking money *in* something is interpreted as investing, and we note that the preposition *in* is appropriate for investing. Risking money; on something is seen as gambling, and we note that the preposition *on* is appropriate for gambling. The example here involving risking something *to*-something is interpreted as exposing, and we note that the preposition *to* is appropriate for exposing.

What we see operating here is a kind of metonymy. Investing, gambling and exposing all contain the notion of risk, so that *risk*, given the appropriate syntactic support, can be used to stand for each of them. Perhaps the ability of *risk* to participate in this metonymy is to be

accounted for by the fact that this verb does not characterize any type of action on its own. It is the type or verb described by Yuri Apresjan as "evaluative" (personal communication): it reveals the evaluated consequences of an action, but it has no other content.

Most of the dictionaries we examined did not identify the three object types, and none of them contained any information, except in the examples they included, dealing with the gerundial complements. And of course none of the dictionaries had any way of relating the various individual senses to a single underlying semantic frame.

Turning briefly to *risk* as a noun, we note that the most frequent uses were as direct object of either the verb *run* or the verb *take*. *Running risks* and *taking risks* have meanings very similar to that of the simple verb *risk*, but they provide the possibility of expressing the evaluation only: since these phrases have no obligatory complement, it is not necessary to include mention of anything specific about the situation, whereas with the verb it is necessary to say something about either the Deed, the Valued Possession, or the Harm. These phrasal expressions welcome *of* + Gerund complements expressing either Deed:

I took the risk of asking my boss for a raise

or Harm:

I took the risk of losing my job

and they also accept a if-clause complement expressing Harm:

Aren't you running the risk that your daughter will never speak to you again?

None of the dictionaries we surveyed told us anything useful about the difference between running and taking risks. Our conclusion was this: that when you speak of *running a risk* you have in mind the situation represented by Figure 1, but when you speak of *taking a risk* you have in mind a situation represented by one of the other two diagrams. Since Figure 1 is included in Figures 2 and 3, there are numerous situations in which either *run a risk* or *take a risk* could be used. In order to test the difference, you need to find a critical sentence which fits one of the diagrams but not the others. Such sentences can be imagined:

The newborn babies in that hospital run the risk of hypothermia.

or

A car parked here runs the risk of getting dented.

In neither of these cases can the version with *take* be used.

We are convinced that our analysis is correct, and that the existence of sentences with *run-risk* which do not allow substitution with *take-risk* supports our understanding of the contrast. However, we found no examples of that type in the actual corpus. And of course, even if we had found such sentences, we would still have to recognize that we cannot find corpus evidence that paraphrasability with *take* is impossible.

The work with *risk* convinced me of the value of a corpus, because, as I said, the simple requirement that we check all of the examples forced us to recognize things that we very probably wouldn't have noticed otherwise. But we could not depend on the corpus alone, since an important judgment that we wanted to be able to make did not receive support from the corpus.

The analysis of the data that we already have has not been completed, but it does in fact seem clear that we need more examples. There are mysteries with the count vs. noncount distinction with this word (a *risk* or *many risks*, vs. *much risk*). We are working with the hypothesis that the noncount form is compatible with *run* but not *take*:

You won't be running much risk if you follow my instructions.

versus

You won't be taking a big risk if you do that.

but not

You won't be taking much risk.

There are also some mysteries having to do with the contexts in which verbs with the different types of complements can occur. It seems that *risk* when accompanied by the Deed complement occurs very often in negative modal form (*I would never risk swimming here* etc.), but we need more examples to see whether this tendency is a real part of the data. In short, I find myself in the end simultaneously convinced (i) that many decisions we have to make about the description of this word cannot be supported by direct corpus evidence, and (ii) that there are decisions that we will be able to make only if we get additional data, from a much larger corpus.

In connection with the next study, I should explain that I have been interested for a long time in words whose grammatical and semantic properties struck me as being completely unique; and *home* is one such word. So when, quite recently, I got my hands on the WSJ section of the DCI corpus (the text of the 1989 *Wall Street Journal*, approximately 8 or 9 million running words), the first thing I did with it was to extract from it all of the sentences containing the word *home*.

Colleagues with access to other corpora who heard about my interest, and who probably worried about the representativeness of my corpus, sent me great quantities of further examples, these taken from the *Grolier's American Academic Encyclopedia* and from on-line newspapers in and around Oxford.

Each of these sources, from written English, produced many hundreds of examples. I work in a third-world university, so except for the London-Lund Corpus, which we bought in better days, we have only corpora that we could get for free. I have access to only relatively small corpora of spoken-language data. I don't think that I will find big surprises when I take a careful look at the conversational data that I have, though there will undoubtedly be big differences in respect to relative frequencies of the usages I've found.

The word *home* has a number of distinguishable uses. Its central use is as a relational noun, seen in phrases like *my home*, *our home*, etc. where it can refer to any place where a person lives, with the resident or residents of the home indicated in a possessive modifier. It is in this central use that the phrase *my home* is to be interpreted as the place where I live, rather than, say, a home which I own.

For interpreting many of the uses of the word we need to appeal to a kind of prototype understanding of this particular cultural unit. A semantic prototype for *home* would probably run something like this:

- a home is a place where people live
- the people who live in the home are members of an intact family
- the home is comfortable and familiar
- each member of the family has unquestioned use of at least some of the objects and facilities in the home
- one lives in the home throughout one's childhood and early youth
- there are many reasons to go away from the home temporarily (shopping, play, travel, education, work, military service, etc.) but after these temporary absences, the natural and expected thing is to return home
- when one reaches the age appropriate for seeking one's fortune, one leaves home and, sooner or later, founds or becomes a part of a new home

A number of lexical and phrasal expressions containing the word *home* appeal to various aspects of such a prototype. A homeless person is someone who has no fixed place to go to after the day's wanderings. *Being homesick* is feeling bad when separated from the familiar and comfortable setting of home and from the people in it. If we remark about somebody that *she left home at age fifteen*, we recognize that this was out of the ordinary. We speak of children of divorced parents as coming from *broken homes*. If I say that I want you, as a guest in my house, *to feel at home*, I am inviting you to treat the objects in my home as objects you can use and enjoy, to relax in the way you would relax in your own home, etc. (We never actually mean just that, of course, but the phrase is intended to give that impression.)

The meaning of *home* that fits the prototype is very closely tied to the notion of family, and in this way *home* differs from *house*. The following contrast shows this distinction quite clearly. If I say that during the first ten years of my life I lived in five different houses, you will assume that my family moved a lot; but if I say that I lived in five different homes, you will assume that I was an orphan and that I lived with five different families, or that I lived in various institutional settings.

In addition to what I spoke of as the central sense of *home*, there are other relational noun usages with meanings that depart in a number of ways from the prototype. With slightly different meanings, two of the other usages can be reflected by their occurrence with the prepositions *to* and *of*; and a third, carrying a considerably different meaning, takes the preposition *for*.

The Barbican Centre provides a permanent home to both the London Symphony Orchestra and the Royal Shakespeare Company.

The African continent was the home of one of the world's oldest civilizations, that of ancient Egypt.

He spent his final years in a home for the aged.

In addition to the necessarily relational uses, *home* also occurs as a plain noun. In this function, not requiring any mention of actual or intended residents, the word is used as a kind of up-market name for *house*. For the noun in this sense, a modifying possessive construction has to identify a relationship other than residence, for example, that of the home's creator. This usage

is said to be an American development, and it is noticeable mainly in the speech of real estate professionals. We see it in sentences like

Our construction company specializes in luxury homes.

Our homes were built with the busy executive in mind.

The focus of my interest in this word has concerned its use without an article, especially when functioning as a locative or directional adverb. Examples of adverbial *home* and its typical contexts are the following:

let's go home

when did you leave home

I just want to stay home

the school principal sent the kids home early today

let's get out there and welcome the troops home

would anybody like to take the leftovers home

I usually work at home

I keep expecting letters from home

I wonder what the folks back home are doing

The adverbial use descends from early dative and accusative case forms, which froze into particular colligations before determiners became popular. The word has both locative and directional adverbial functions, at least in American English. It occurs with the prepositions *at* and *from*, but not *to* when it is a complement of a verb. (That is, we can say *go home* but not *go to home*. However, in structures in which *to* is independently required, the combination of *to* + *home* is possible: *from work to home*, *close to home*, etc.)

The adverb behaves - most clearly in the case of its occurrence with transitive verbs - like a verbal particle. That is, as with other particles like *off*, *away*, etc., we find alternations between Object + Particle and Particle + Object orders.

Would anybody like to take home the leftovers?

Would anybody like to take the leftovers home?

There are numerous reasons for my interest in adverbial uses of *home*. One is that they present a problem in cohesion semantics: in the case of the noun *home*, the "resident" is identified by a possessive determiner, but in the adverbial use we have to figure out who lives in the house from the context. This fact is indirectly revealed in definitions of the word *home* through the use of the word *one* - and one of my interests in lexicographic traditions is the conventions for using the word *one* in defining phrases.

In the *Concise Oxford Dictionary* (1990) we find *home* defined as "the place where one lives"; in the *Chambers Twentieth Century Dictionary* (1983) it is "the residence of one's family"; in the *Collins English Dictionary* (1986), "the place or a place where one lives"; in *Webster's*

Third New International Dictionary (1986), "one's principal place of residence"; in the *Random House Dictionary* (1987), "the place in which one's domestic affections are centered"; and so on. This definitional pattern distinguishes *home* from *house*, where the definers never use the word *one*, but are more likely to speak of something like "a structure in which people live "or " a building used as a home.

Because of the connection between the adverbial uses and the central sense of the noun, we can think of the prepositionless noun as meaning "the place where one lives", the locative adverb as "at the place where one lives", and the directional adverb as "to the place where one lives".

The felt appropriateness of the word *one* in these definitions reveals an anaphoric element in the meaning of the word. One part of the process of giving a semantic interpretation to expressions containing the adverb *home*, then, is that of establishing the cohesive link between this hidden anaphoric element and some other part of the text which can provide its antecedent. When I go home it's to my home; when the factory boss sends the workers home, it's to their homes, etc. One of my interests was in figuring out whether there are any strict principles determining what controls or binds the hidden anaphor in *home* and whether what we know about the anaphoric properties of *one* allows us to use it in formulating the definitions of adverbial *home* in a way which predicts which cohesive links are possible and which are not. A second interesting fact about adverbial *home* is its participation in multiple contrast sets. One of the discussions of *home* in Quirk et al. (1985) points out the quasi-antonymy relation the word has with *abroad* and *out*:

We were *abroad* during the last few summers, but this year we're staying *at home*.

I've gone *out* the last few nights, but tonight I'm staying *at home*.

One question I'd like to ask is whether we are dealing with clearly distinguishable contrast sets here. The adjective *short* has very similar meanings in the contrast set in which it is opposed to *long* and in the one in which it is opposed to *tall*, but it is quite clear that it has separate if related senses precisely because of its participation in these two antonymy relations. What can we say about *home* in this respect?

A third point of interest relates to certain differences between American English and British English. The usage notes in some dictionaries **tell** us that in British English *be home* is used only to refer to a situation in which someone has freshly arrived from elsewhere. I wanted to see if there are any traces of this distinction in the American English examples.

A fourth reason for being interested in *home* relates to my interest in deixis. Twenty years ago, as a part of a series of lectures on deixis, I read a paper in which I presented what I then believed to be a true account of the; English verbs *come* and *go* (see Fillmore 1975). The connection with deixis] is that in describing *come*, one has to say something about the presupposed location of one or both of the speech-act participants at the destination of the journey. Independently of that deictic feature, I claimed in my paper that a temporal adverb associated with a *come* expression identified the arrival time, whereas a temporal adverb

adjoined to a *go* expression identified the departure time. To see what I mean, imagine Max at a late night party and people talking about his return home after the party. The sentence

Max went home at midnight.

would be interpreted as telling us what time Max left the party, but

Max came home at three in the morning.

said by somebody in his home, would inform us of the time at which he arrived at the house. The generalization I proposed is that a time-phrase with *go* indicates the departure time, a time-phrase with *come* indicates the arrival time.

I believed, then, that the interpretation differences I reported had to be described as a difference in the semantic structures of *go* and *come*. If anybody had asked me about it, I surely would have said that the fact that I used the word *home* in my examples to indicate the destination of the journey was purely accidental. Anything else would have done just as well, I would have said.

Once in a while, in the intervening years, I worried about the fact that a sentence like

He went to the dentist's at two o'clock.

doesn't really mean the same thing as

He left for the dentist's at two o'clock.

as my generalization would have predicted, but I tended to think that there must be some special problem with such sentences. I now believe, as you have guessed, that the difference has a lot to do with the word *home*.

There was still another reason for my interest in *home*. I have a general cross-linguistic interest in the concept of the "home base" as a feature in lexical semantic systems.

A home base feature is present in the semantic systems of many languages, and sometimes the home base category interacts with or contrasts with the other deictic categories in the verbs of motion. I know of several such systems in the native languages of the Americas, but the phenomenon might well be much more widespread than that.

In Japanese the idea of going home, coming home, returning home, getting home, etc., is expressed using the verb *kaeru*, which is usually translated as "return". And the idea of sending somebody home, bringing or taking somebody home, is expressed with the causative form of *kaeru*, namely *kaeraseru*. These verbs usually occur in construction with some secondary verb indicating the difference between coming and going, or that between sending and accompanying.

I have always thought that the Japanese verb *kaeru* really means what the English verb *return* means, but that it is simply conventionally used in the context of talking about going home. This seems reasonable, since every journey to one's home is an instance of returning. I now think, however, that it really means "to go home" and that its use in some contexts with temporary starting places is a separate development. The difference can be seen in talking to Japanese or foreigners about going back to Japan. If I want to ask a Japanese person

when he's going back to Japan, I can use the word *kaeru*, because he is going home. But if I am talking to a foreigner who may have visited Japan many times to ask him if he is ever planning to return to Japan, I cannot use the verb *kaeru*. I have to say something that means "go again". Japan is not that person's home, and so *kaeru* is not appropriate. The idea that *kaeru* and *return* mean the same thing is not only a mistake made by English-speaking people who are learning Japanese. It works in the other direction too. I recently learned about an anti-American demonstration in Japan in which the protestors carried placards in English urging Americans to return. The addressees of this message might have found it quite friendly and welcoming, were it not for the accompanying shouting and clenching of fists.

The WSJ corpus yielded about 450 sentences with determiner-less instances of *home*, and I'll briefly survey the collection now. The examples sort themselves into literal and figurative, and I begin with the literal. Of these, the examples can be divided into those expressing location, those expressing going away from the home, and those expressing returning to the home. The location examples can express location at the home or location away from the home. The examples of returning to the home are further divided into those that express arrival at the home, those that express setting out on the homeward journey, and those that express transit. Superimposed on these path differences is the distinction between intransitive and transitive, i.e. between plain and caused movement.

One group of the location examples simply described things that were at the resident's home, expressed as objects of the verb *have* or objects of the preposition *with*.

He believes every family should have a Bible at home.

According to the poll, 19% of respondents already have computers at home.

It would be nice if each of us had a wife at home to anticipate and meet our needs,...

I remembered being fired at age 44, with five children at home.

Sentences about the resident not leaving home used the verb-phrase *stay at home* and *stay home*. The latter possibility apparently does not exist in British English.

Mothers who work should be subsidized more than those who stay at home.

I don't have the personality to stay at home.

If I stayed at home, I'd be looking at the walls. Friends criticized her for not staying at home.

Then maybe I could stay home and have seven children and watch Oprah Winfrey.

I think a lot of people got scared and stayed home.

ABC wanted comedies that would appeal to the kind of people who stay home Saturdays.

Kaye Myers challenges the view ... that men should work and married women should stay home with the children.

Stay specifically communicates the idea of not going out or away, but we also have expressions with *be*, simply indicating the fact of being at home. Both languages accept *at*

home.

The company's chairman ... and another top official were at home yesterday.
Subscribers won't need to be at home during the day for in-home service calls.

There is a difference between the two languages when *be* is followed by *home* without a preposition. The usage notes tell us that there is no distinction in American English between *be home* and *be at home*, but that in British English *be home* is used only to express the idea of having freshly returned from somewhere. Another way to think about this is to say that in British English, *home* without a preposition is only a dynamic or directional adverb, never a static or purely locational adverb. Thus perhaps *he is home* has a structure that is a bit like *the ball is over the fence*, where we are indicating the location of something by saying that it just got there.

The following examples would presumably not be accepted by British speakers.

Japanese tradition says she should be home taking care of her two preschool children.

Because of her inability to be home to care for a kitten, she was counseled instead to adopt a cat.

"At least she can die knowing she is home."

For the fresh-return sense, American English also prefers the preposition-less form.

"Mona, I'm home!"

"Will he be home in time for dinner?"

He barely had time to tell the news media "I am happy to be back I home" before one of his bodyguards tugged his elbow and said)

"Comrade, let's move".

If a speaker of American English comes home, sees nobody in the house, wants the people in the house to know that he is back; he will shout *I'm home*, and surely not *I'm at home*.

I am not sure, but I think we can summarize the difference by saying; that in British English the prepositionless form requires the "fresh arrival" meaning, and in American English the "fresh arrival" meaning requires the prepositionless form.

There is another difference between *be home* and *be at home* in American; English. The form without the preposition can only express the resident's location, the form with *at home* can express the location of some possession of the resident. Thus, it is possible for me to say that *my computer is at home* but I can't say that *my computer is home*.

There are expressions talking about people not being at home. The most common phrase is *away from home*.

The children were away from home for 16 days altogether.

The ASPCA doesn't give young kittens, which are more in demand than cats, to

people who are away from home during the day.

School administrators walk a tightrope between the demands of the community and the realities of how children really act when they are away from home.

The word *away* is interesting in this respect. Being *away from home* can signal a short absence or a long absence, but simply being *away* suggests an absence of at least one night. In this it contrasts with being *out*. Thus if somebody calls for my wife asking if she is at home - or, in America, if she is home - I could answer that she is away, the assumption being that she won't be back until at least tomorrow, or I could answer that she is out, the assumption being that she will return in the same day. The examples I mentioned from Quirk et al. (1985) earlier contrasted *at home vs. abroad* and *at home vs. out*, but I think that *away* is a third alternative.

When you are away from home you can still communicate with your family. The "directionality" with the adverb *home* can also be that of a communicating act, as shown in two of the WSJ examples

You can always call home if you're lonely.

In the spring of '40 they stopped writing home.

Another attested sentence, absent from the WSJ corpus, is *E.T., phone home*, illustrating the same point.

Leaving home can be for a short period or a long period. One can go shopping, one can go on a trip, or one can leave home for good. Examples of each type were found in the corpus:

Many residents leave home without locking doors.

Most travelers are leaving home for fun rather than business.

If you'd rather have a Buick, don't leave home without the American Express card.

Miss Johns ran away from home to California at age 14, got a job as a bank teller, earned a high school equivalency degree and became a trading assistant at Drexel Burnham Lambert.

At an early age, Wonda left home and married.

Verbs indicating arrival at home include *arrive*, *get*, *return*, and *come*.

Warner arrived home to tidy the house and prepare a nourishing meal for the brats.

Four months after he got home, he and his wife separated.

"I phoned my wife from there. 'Put on the coffee', I said, 'I'm coming home for good'."

"You can come home from work at 6 o'clock, and they call it 'abandonment'."

Verbs indicating "going home" included simple *go*, the general directional verb *head* (*heading home* is going in the direction of home), and manner and means verbs like *hurry*, *run*, *drive*, *ride a bike*, etc. We find that *home* is also a possible complement for nouns designating journeys:

On the flight home she kept worrying about the children.
The journey home was to take three days.

The transitive verbs we find in the corpus include *bring*, *take*, and a number of verbs suggesting the idea of carrying, and *send*, *order*, *summon*, etc., suggesting the idea of giving orders.

The remaining examples were metaphorical, with *at home* used in the meaning of being competent (*The pianist was at home with Chopin*); *pounding*, *driving*, *hammering*, *nailing or pressing a point home*, meaning something like "try energetically to convince"; *hit home*, *strike home*, etc., meaning "to affect one deeply".

I turn now to the various semantic problems we left hanging. One was the question of interpreting the anaphoric element in *home*, and the appropriateness of the word *one* in the definition of *home*. Dictionary entries for idioms with variable possessive pronouns distinguish two types, along the lines of *to blow one's nose* and *to pull someone's leg*. The possessor, in the case of the idioms listed with *one's* is always the subject of the verb (*I blow my nose*, *you blow yours*), but in the case of the idioms listed with *someone's* it is distinct from the subject (*I'm pulling your leg*, **I'm pulling my leg*).

I will allow myself to use the word "control" to express the relation between the antecedent and the anaphoric element in *home*. In all of the examples of intransitive verbs in the corpus the controller was the subject. With the transitive verbs the controller was the subject with some verbs - for example *bring*, *carry*, *tote*, and *take* - but the object with others - for example *send*, *summon*, *order*, etc. In the case of *be home* or *be at home* it was always the subject. However, this is a case where the corpus has let us down, it failed to show that the subject with *at home* wouldn't have to be the resident of the *home* (since I can say that I left my computer at *home*), and it failed to show that with *bring* and *take* the object could be the controller. (Actually there was one example with *take*, concerning a limousine that was waiting to take some judges *home*.) It is certainly possible to say things like

A policeman brought my husband home last night.

It is also possible for *send* to be used with the subject, not the object, controlling the anaphoric element of *home*, as in

If, when I travel, I buy books, I always send them home rather than carry them home.

But there's still more to say. In the case of both the intransitive verbs and the transitive verbs, there are contexts in which the resident of the home can be introduced with a *with* phrase:

My dog is so friendly he 'll go home with anyone. And we can imagine a sentence with three possible controllers of the anaphoric element of *home*. Consider the sentence *The teacher sent Jimmy home with Mary.* It could mean that she sent the kids to her own home; it could mean that she sent Jimmy to his home, in Mary's company; or it could mean that she sent Jimmy to Mary's home (in Mary's company). The context for this last case might be that Jimmy's parents had had a family emergency and had to leave town, and that Mary's parents had agreed to take care of Jimmy during their absence. *The teacher sent Jimmy home with Mary,* then, would mean that it was to Mary's home.

It seems clear that the relationship between the anaphoric element of *home* and its antecedent is not controlled syntactically but semantically. The controller can always be the subject, but otherwise you simply have to know who the travelers are. Furthermore, paraphrasing *home* with "at one's home" or "to one's home" will not make it possible to identify the antecedent of the anaphoric element of *home*, because the referential anaphor *one* is always bound to the clause's subject.

You have undoubtedly noticed that I resorted to made-up examples and imagined contexts. But that, of course, is because I'm pretty sure that what I am claiming about the fact of the matter is right, but the corpus didn't give any evidence on it one way or the other.

I turn now to the paradigmatic semantics of *home*. A part of knowing what function a word performs in a given context is knowing what it is being used "in contrast with" in that context. Recall what we have already noticed about the ability of *at home* to be in contrast with *out*, *away* or *abroad*.

In the WSJ corpus a great many of the sentences contained explicit indications of the alternative to *at home*, either in a way in which the contrast was presented directly (*at home and abroad*, *at home or at work*, *in school but not at home*), or in some less direct way. In fact, in just this respect it is clear that in the case of *home* we need more than a single sentence to learn the facts, since we had some sentences like *At home, however, things were different.* To understand the scale of the intended contrast for this sentence we have to know what the preceding sentence was.

These contrasts show us that the English category of *home* involves considerable variation of scale.

Advertisers claim virtues for their products overseas that they are forbidden by law to claim at home.

Senators and representatives don't always say the same thing in Washington that they say at home.

The yen is powerful overseas but has little purchasing power at home.

Parents wonder if their children behave better at school than they behave at home.

These children speak one language at school and another language at home.

In many cases the contrast was covert, discoverable only by figuring out why it seemed relevant to use the adverb *at home*. It is relevant to say that some parents teach their children at home, since the usual thing is to have children educated at school. It makes sense to talk about

people who shop at home with a computer, or that it bad times people tend to eat at home, because these activities are understood as things that one could carry out in shops or restaurants.

A Bulgarian linguist, Svillen Stanchev, who has been visiting the Berkeley campus this year, went through my WSJ examples, and concluded that the English word *home* was a translator's nightmare. Most of the sentences could not be translated using a Bulgarian equivalent of *home*. There seem to be three reasons for this. One is that Bulgarian does not have the scale variations that English has, which allows us to use *at home* to refer to being in one's house, neighborhood, town, state, country, or planet. A second is that Bulgarian doesn't seem to allow the distributive interpretation of adverbial *home* with multiple travelers. In a sentence about the director of a factory sending the workers home after an accident, a translation with *home* would suggest that they all Went to the same home, but a translation that made explicit that each worker went to his own home sounded silly, since the point was that the workers had to leave the factory and in the context there was no reason to make sure that each one ended up in his own proper home. The third reason – and this was the biggest surprise I got from the corpus - is that many English expressions with *home* appear to be simply negations of the other member of the contrast set. That is, to say that *Joe is at home* sometimes means simply that Joe is not at the other place where he might instead relevantly be.

It appears that with the word *home* a potential three-way alternation has come to be seen as a two-way contrast. If on a given day I could be expected to be either at home or at work, that sounds like a two-way contrast, but we know, of course, that in reality there is a third possibility: I might go someplace else altogether - for example, to the beach, to a coffeehouse, or what have you. But we find in the corpus lots of expressions about *being at home* or *going home* or *being sent home* that give the impression of only a two-way contrast. We read that *in a bad business climate customers stay home in droves*. (There aren't many homes large enough for people to assemble in droves.) Because of an accident the factory workers stayed home. One of the most striking examples of this that I have noticed - not from the WSJ corpus but from a recent news report - was a sentence which described non-voters as people, who stay home on election day. Since elections in the United States are traditionally held on Tuesdays, when most people work, it is not at all likely that the people who didn't vote stayed home.

It is this use that I now associate with the "departure time" interpretation of *go home*, since *go home* has to be interpreted as going away from the place where one is. Mr Stanchev tells me that the slogan YANKEE GO HOME would not work in direct Bulgarian translation, since it focuses on the wrong end of the journey.

A careful study of sentences with *risk* and with *home* has revealed facts about the uses and meanings of these words that have not been well described in existing grammars or dictionaries, and has given me reasons to be absolutely committed to the use of corpus evidence. But it is also true that in thinking through the consequences of the various hypotheses that observed corpus data evoked, other judgments needed to be brought in. Atkins and I think that we understand the difference between *run a risk* and *take a risk*, but we didn't find the critical examples in the corpus. But even if we had found sentences which worked with *run* but which wouldn't have tolerated replacement of *run* with *take*, we still have to face the reality that there are no corpora of starred examples: a corpus cannot tell us what is not possible. The cohesion problem with *home* is not syntactically resolvable, but almost all of the examples in the actual corpus did suggest that antecedents could be found in the subjects

or the objects. The possibility that they could also be found in the objects of the preposition *with* was not shown in the corpus, and this seems to be an accidental gap.

As I said at the beginning, my concern with corpora is with the possibility of amassing enough examples to cover a particular domain more thoroughly than armchair linguist could possibly manage without this sort of help. So one kind of corpus linguist should find this encouraging: there are really good reasons for building corpora, and as far as I'm concerned the bigger the better. But what I have been saying is probably not encouraging to people who want to do most of their analysis without expecting anyone to have to sit down and stare at the examples one at a time to try to work out just what is the intended cognitive experience of the interpreter, what are the interactional intentions of the writer, and so on. Should, it ever come about that linguistics can be carried out without the intervention and suffering of a native speaker analyst, I will probably lose interest in the enterprise.

Notes

1. COBUILD: Collins Birmingham University International Language Database; also a metonymic name for the *Collins COBUILD English Language Dictionary*, 1987 (Editor-in-Chief: John Sinclair).
2. This is part of the DCI (Data Collection Initiative) corpus of the Association for Corpus National Linguistics; segments of the corpus were provided to the University of California at Berkeley through the courtesy of Mark Liberman of the University of Pennsylvania. (I am grateful to the Institute of Cognitive Studies for providing an electronic home for our campus's growing collection of linguistic corpora as well as facilities for accessing and processing them.)
3. The corpus was provided by the Japanese Telecommunications firm NTT, in connection with an NTT-sponsored research project which I am directing.
4. Two studies based on this work are soon to appear: Fillmore - Atkins forthcoming a, b. The summary presented here repeats material found in those articles.
5. A discussion of this approach can be found in Fillmore 1985.
6. For a representative work on decision theory, which uses the notation, see Raiffa 1970.

References

Fillmore, Charles J.

1972 "A grammarian looks to sociolinguistics", *Georgetown University Monographs on Language and Linguistics* 25: 275-287. Washington, DC: Georgetown University Press.

1975 "Santa Cruz lectures on deixis 1971", reproduced by the Indiana University Linguistics Club; the lecture "Coming and going", under the title "How to know whether you're coming or going", was revised and reprinted in Karl Hyldgaard-Jensen (ed.), *Linguistik 1971*: 369-379. Athenaeum.

1985 "Frames and the semantics of understanding" *Quaderni di Semantica* 6: 222-254.

Fillmore, Charles J. - Beryl T.S. Atkins

Forthcoming a "Toward a frame-based lexicon: the semantics of RISK and its neighbors", in Adrienne Lehrer - Eva Kittay (eds.), *Frames, fields, and contrasts*. Hillsdale, NJ: Lawrence Erlbaum.

Forthcoming b

"Starting where the dictionaries stop: the challenge of corpus lexicography", in:
B.T. Atkins - A. Zampolli (Eds.). Computational approaches to the lexicon. Oxford:
Oxford University Press.

Polanyi, Michael

1958 Personal knowledge: towards a post-critical philosophy. Chicago, Illinois:
University of Chicago Press. Quirk, Randolph - Sidney Greenbaum - Geoffrey Leech -
Jan Svartvik

1985 A comprehensive grammar of the English language. London:
Longman.

Raiffa, Howard

1970 Decision analysis: introductory lectures on choices under
uncertainty. Reading, MA: Addison Wesley.