THE PHRASAL LEXICON

Joseph D. Becker

ABSTRACT

Theoretical linguists have in recent years concentrated their attention on the productive aspect of language, wherein utterances are formed combinatorically from units the size of words or smaller. This paper will focus on the contrary aspect of language, wherein utterances are formed by repetition, modification, and concatenation of previously-known phrases consisting of more than one word. I suspect that we speak mostly by stitching together swatches of text that we have heard before; productive processes have the secondary role of adapting the old phrases to the new situation. The advantage of this point of view is that it has the potential to account for the observed linguistic behavior of native speakers, rather than discounting their actual behavior as irrelevant to their language. In particular, this point of view allows us to concede that most utterances are produced in stereotyped social situations, where the communicative and ritualistic functions of language demand not novelty, but rather an appropriate combination of formulas, cliches, idioms, allusions, slogans, and so forth. Language must have originated in such constrained social contexts, and they are still the predominant arena for language production. Therefore an understanding of the use of phrases is basic to the understanding of language as a whole.

You are currently reading a much-abridged version of a paper that will be published elsewhere later.

It's Not WHAT You Say...

Like all other scientists, linguists wish they were physicists. They dream of performing classic feats like dropping grapefruits off the Leaning Tower of Pisa, of stunning the world with pithy truths like "F=ma", and in general of having language behave in an orderly way so that they could discover the Universal Laws behind it all. Linguists have a problem because language just ain't like that. Physical laws are very basic, general-purpose constitutents of the universe, so the Creator was forced to keep them elegant and potently simple. Language, by contrast, was recently invented by Man for the sole purpose of giving his Fellow Man the low-down; for this reason language is inextricably bound to humans, human communication, and the circumstances of human communication.

Nevertheless, linguists have tried to extricate the inextricable. The first level of extrication is to declare that there is such a thing as "language", and that all other items like human beings, psychology, physical objects, events, and social situations are "non-linguistic" and should please go away. The second level of extrication is to declare that there is such a thing as "competence", which is what a language would be like if a decent mathematician had drafted it, and that all other items like everything that people actually say, write, or think are "performance" and should please go away.

And so the "modern" linguist spends his or her time starring or unstarring terse unlikely sentences like "John, Bill and Tom killed each other" (to pick one at random from a recent journal), which seethe with repressed frustration and are difficult to work into a conversation. These example sentences bear no discernable resemblance to the sentences which compose the text that purportedly explains them -- yet the linguist's own sentences are also alleged (implicitly) to be drawn from the same English Language! Perhaps it is time that editors, or at least readers, began applying Becker's Criterion to their readings in linguistic theory:

> BECKER'S CRITERION
>
> Any theory (or partial theory) of the English Language that is expounded in the English Language must account for (or at least apply to) the text of its own exposition.

Using this handy guideline, you can pretty much wipe your theoretical linguistics shelf clean and start over.

This is not to say that there has been no constructive value in the attempt to physicize (i.e., mathematicize) natural languages. Without question, many important regularities have been discovered thanks to this approach. Nevertheless, I feel that for a science to attempt to deny the existence of nearly all of its subject matter (English as she is spoke) and to deny the existence of the substrate of its subject matter (communication between human beings) is (1) scientifically dishonest, and therefore (2) ultimately self-defeating. Modern theoretical linguistics is rather clearly self-defeated already, and linguists have yielded the scientific initiatve to untutored computer types. The latter are relatively unskilled in sophistry, but they have been trained to unflinchingly confront large-scale, complex, inelegant, real-world behavioral systems such as language, and to attempt to understand the workings of these systems without vainly pretending that they can be reduced to pristine-pure mathematical formulations.

Let's Face Facts

Most of this paper will be devoted to organizing and presenting the facts concerning our knowledge of English phrases, rather than to expounding a theory that would explain these facts. There are four reasons for this. First, the conspiracy of silence which has surrounded these facts has lasted for so many years that it is worthwhile to take a good look at them. Second, I hope to present the beginnings of a taxonomy for lexical phrases. Third, the phrases themselves are more interesting than any theory will ever be. And of course fourth, I don't have a theory. But actually the elements of such a theory are already clear and present, as I will suggest in the last two sections.

The structures I am discussing in this paper have often been swept under the rug by means of the disclaimer "Oh, that's an idiom." The result is that we now find a lot of apples and oranges under the same rug. In order to replace the useless term "idiom" with a taxonomy of some substance, I am proposing six major categories of lexical phrases. The categories are listed in order of increasing "size" of the members, as the descriptions below should make apparent.

CLASS I: Polywords

Nature: Multi-word phrases admitting no variability, interchangeable with single words or concepts.

Function: The same as single words. Euphemisms sometimes fall into this category.

Examples: the oldest profession (n.) [= prostitution]
to blow up (vi., vt.) [= to explode]
for good (adv.) [= forever]

CLASS II: Phrasal Constraints

Nature: Units consisting of a small number of words, some of which constrain the variability of others; in the limiting case the whole phrase is invariable.

Function: Often specify how a particular expressive function is to be applied to particular semantic material.

Example: If we wish to say that something happened coincidentally, and we wish to underscore that assertion, we say that it happened "by pure coincidence"; stronger yet is to say "by sheer coincidence."

CLASS III: Deictic Locutions

Nature: Phrases with low variability, short-to-medium length.

Function: Serve as clauses or whole utterances whose purpose is to direct the course of conversation, i.e. the flow of expectations, emotions, attitudes, etc.

Examples: for that matter
[= "I just thought of a better way of making my point"]
...,that's all. [= don't get flustered]

CLASS IV: Sentence Builders

Nature: Phrases up to sentence length, often containing slots for "parameters" or "arguments".

Function: Provide the skeleton for the expression of an entire idea.

Example: (person A) gave (person B) a (long) song and dance about (a topic).
[= "A tried to convince B of something, and was cynical and perhaps less than truthful about what he said"]

CLASS V: Situational Utterances

Nature: Usually complete sentences, little variability.

Function: Utterances which are known to be the appropriate thing to say in certain circumstances; may be used out of context for effect.

Examples: How can I ever repay you?
[expresses moderate-to-large gratitude in response to some kindness]
It only hurts when I laugh!
[expresses the unimportance of some apparent affliction; originated as a Vaudeville joke]

CLASS VI: Verbatim Texts

Nature: Texts of any length memorized verbatim, or approximately so.

Function: Used as substance for quotation, allusion, variation, and occasionally direct usage.

Examples: Better late than never [proverb]
How ya gonna keep 'em down on the farm?

[song title]
99 and 44/100 percent pure
[advertising slogan]

The Region of Extraction and Processing

In the textbook from which I first studied Russian, each lesson began with a section entitled "Idioms and Common Expressions", where they put everything the student couldn't understand on the basis of his current vocabulary. I'm sure I learned all of these phrases, but three in particular stand out in my memory: Vyera Aleksyeevna otkrivayet dyver' (Vera Aleksyeevna opens the door), vsya chisto literaturnaya dyeyatel'nost' prekonchalas' (all purely literary activity came to an end), and raion dobichi i obrabotki (the region of extraction and processing). Sad to say, this book did not prepare me for my visit to the Soviet Union. I discovered that the book's expression for asking the time of day was not the one that people used, and that I had no idea whatsoever how to get a 10-kopek piece changed into three 3's and a 1 so I could use the vending machines.

This is the reality of language: In order to survive in society we've got to know what to say, and we usually know it in advance by memorizing it. The suggestion here is that the wonderful feats of the human intellect, such as the use of language, are based at least as much on memorization as on any impromptu problem-solving (in this case, the generation of novel utterances). We owe this insight to Minsky, his), and I am glad to see it finally being recognized.

What does all this imply for a theory of language production? It implies to me that the process of speaking is Compositional: We start with the information we wish to convey and the attitudes toward that information that we wish to express or evoke, and we haul out of our phrasal lexicon some patterns that can provide the major elements of this expression. Then the problem is to stitch these phrases together into something roughly grammatical, to fill in the blanks with the particulars of the case at hand, to modify the phrases if need be, and if all else fails to generate phrases from scratch to smooth over the transitions or fill in any remaining conceptual holes.

My guess is that phrase-adaption and generative gap-filling are very roughly equally important in language production, as measured in processing time spent on each, or in constitutents arising from each. One way of making such an intuitive estimate is simply to listen to what people actually say when they speak. An independent way of gauging the importance of the phrasal lexicon is to determine its size.

There is no dictionary in English that comes even close to encompassing the variety of lexical phrases discussed in this paper, from cliches to sentence patterns to wisecracks to song lyrics. The most respectable phraseological dictionary I have found is an English-Russian one which I bought when I found it to contain knee-high to a grasshopper. This book has 25,000 entries. I estimate that I recognize about half of them (the rest being British or archaic), and that about half of the ones I think of are not in the book. This gives us a ballpark estimate of around 25,000 phrases, which is very much the same magnitude as our single-word vocabularies. And this does not include most of the Class VI Verbatim Texts.

All in all, we must conclude that the phrasal lexicon is very real. Even excluding long verbatim texts, we probably know as many or more whole phrases than we know single words (and I suspect the disparity would be even greater for the under-educated, i.e. almost all of humanity, since book-learning adds more words but few social situations to the individual's experience).

Because lexical phrases are real, they have an advantage over transformations and other such chimeras in that they are actually observable. Having read this paper, you will have no trouble hearing the cliches as they come tripping off the tongues of the folks that surround you. And, for better or for worse, you will feel them popping out of your own brain when you speak and when you write. This experience should give you a better understanding of the process of language production than any theory I could espouse to you on paper.

End Test

Which brings us directly to the final and ultimate question: Does Becker's paper meet Becker's Criterion? Does the view of language propounded here at least apply to the language in which it is propounded? Well of course:

Some Lexiccal Phrases Encountered in This Paper

concentrate (one's) attention on
to give (a person) the low-down
inextricably bound to
(verb) the un(verb)able
to work (something) into a conversation
it is time that
to start over
this is not to say that
English as she is spoke
conspiracy of silence
clear and present
for the most part
an integral part
to sweep under the rug
apples and oranges
as (something) should make apparent
in the limiting case
out of context
sad to say
the time of day
in advance
What does this imply for...?
the case at hand
if need be

if all else fails
my guess is that
as measured in
as far as I know
no (n.) comes even close to
a ballpark estimate
(very) much the same
all in all
we must conclude that
to have (an/the) advantage over
tripping off the tongue
for better or for worse
which brings us to

Yet if we look carefully at the text of this paper, we find fairly long stretches without any apparent lexical phrases. About these, three points should be made.

First, most of the lexical phrases that we actually use in speaking or writing are so humble and uninteresting that they would never appear on a list devoted to picturesque expressions like Davy Jones's Locker. Yet these humble patterns do most of the work of language production for us. For example, the first sentence of this paper (come to think of it) is composed of three humble patterns:

Like all other scientists, linguistics wish they were physicists.

A: Like (pl. n.),...
B: all other (pl. n.)
C: (person) wishes (he/she) were (something)

You may say that these patterns are implicit in the lexical entries for the individual words plus the grammar of English, but I say that I am also familiar with the patterns themselves. To write the first sentence of this paper, I brought forth pattern C to express my main thought, then took pattern A and nested pattern B into it to express my subordinate thought, and tacked this onto the front of pattern C. No wonder my paper conforms to my own theories!

Second, nothing in this paper says that so-called "generative" processes do not play an important role in language production. I assert that their role is equal to or less than that of phrasal processes, but that does not make it zero.

Third, writing is a specialized skill that is not identical to speaking, and technical writing is especially so. It takes us years of strenuous effort to learn to write, beginning long after we have completed learning to speak, and most of us never learn to write very well at that. I make this point, even though it somewhat weakens the impact of Becker's Criterion, in order to counteract the tendancy of intellectuals to believe that their language is typical of the language as a whole. It isn't. In particular, narrative monologues do occur in the conversations of the unenlightened, and I think you will find them to be much more phrase-based than is any technical essay.

And so I conclude that the rather messy taxonomy given in this paper, and the messy Compositional notion of language production, have a fair amount of truth to them when we look at what people actually say, think and write. Indeed, I suggest that the realer the text, the messier and truer these notions become. All of this can of course be summed up in a single elegant principle, namely:

BECKER'S RAZOR
Elegance and truth are inversely related.

Put that in your phrasal lexicon, and invoke it!