Lexis

In linguistics, lexis describes the storage of language in our mental Lexicon as prefabricated patterns that can be recalled and sorted into meaningful speech and writing. Recent research in corpus linguistics suggests that the long-held dichotomy between grammar and vocabulary does not exist. Lexis as a concept differs from the traditional paradigm of grammar in that it defines probable language use, not possible language usage. This notion contrasts starkly with the Chomskian proposition of a "Universal Grammar" as the prime mover for language; grammar still plays an integral role in lexis, of course, but it is the result of accumulated lexis, not its generator. In short, the Lexicon is • Formulaic: it relies on partially-fixed expressions and highly probable word combinations • Idiomatic: it follows conventions and patterns for usage • Metaphoric: concepts such as time and money, business and sex, systems and water all share a large portion of the same vocabulary • Grammatical: it uses rules based on sampling of the Lexicon • Register-specific: it uses the same word differently and/or less frequently in different contexts

1. Formulaic Language

In recent years, the compilation of language databases using real samples from speech and writing has enabled researchers to take a fresh look at the composition of languages. Among other things, statistical research methods offer reliable insight into the ways in which words interact. The most interesting findings have taken place in the dichotomy between language use (how language is used) and language usage (how language could be used).

Language use shows which occurrences of words and their partners are most probable. The major finding of this research is that language users rely to a very high extent on ready-made language "chunks", which can be easily combined to form sentences. This eliminates the need for the speaker to analyze each sentence grammatically, yet deals with a situation effectively. Typical examples include "I see what you mean" or "Could you please hand me the …" or "Recent research shows that…"

Language usage, on the other hand, is what takes place when the ready-made chunks do not fulfill the speaker's immediate needs; in other words, a new sentence is about to be formed and must be analyzed for correctness. Grammar rules have been internalized by native speakers, allowing them to determine the viability of new sentences. Language usage might be defined as a fall-back position when all other options have been exhausted.

2. Context and Co-Text

When analyzing the structure of language statistically, a useful place to start is with high frequency context words, or so-called Key Words in Context (KWICs). After millions of samples of spoken and written language have been stored in a database, these KWICs can be sorted and analyzed for their co-text, or words which commonly co-occur with them. Valuable principles with which KWICs can be analyzed include:

• **Collocation:** words and their co-occurrences (examples include "fulfill needs" and "fall-back position")

• **Semantic Prosody:** the connotation words carry ("pay attention" can be neutral or remonstrative, as when a teacher says to a pupil: "Pay attention!" (or else)

• **Colligation:** the grammar words use (while "I hope that suits you" sounds natural, "I hope that you are suited by that" does not).

• **Register:** the text style a word is used in ("President vows to support allies" is most likely found in news headlines, whereas "vows" in speech most likely refer to "marriages"; in speech, the verb "vow" is most likely used as "promise").

(partially adapted from Lewis, 1997)

Once data has been collected, it can be sorted to determine the probability of co-occurences. One common and well-known way is with a concordance: the KWIC is centered and shown with dozens of examples of it in use, as with the example for "possibility" below.

Concordance for POSSIBILITY

```
  bout to be put on looks a real possibility. Now that Benn is no longer Hiett, says
 that remains a real possibility: As part of the PLO, the PLF Graham added. That's a
 possibility as well," Whitlock admitted. Severe pain was always a possibility. Early
 in the century, both that, when possible, every other possibility, including speeches
 by outside that we can, that we use every possibility, including every possibility of
 could be let separately. Another possibility is `constructive vandalism' a people
 reject violence and the possibility of violence can the possibility the French vote
 and now enjoy the possibility of winning two seats in the immediately investigate the
 possibility of criminal charges and that her Sri Lankan sources say that the
 possibility of negotiating with the Tamil Sheikhdoms too there might be the
 possibility of encouraging agitation. the twelve member states on the possibility of
 their threatening to Marie had already looked into the possibility of persuading the
 [f] a function of dependency, but the possibility of capitalist development, were
 almost defenceless. The possibility of an invasion had been apparent oddly and are
 worried about the possibility of drug use, say so. Tell them was first convened to
 discuss the possibility of a coup d'etat to return the in the mi5 line and in the
 possibility of the state being used to smear reasons behind the move was the
 possibility of a new market. Cheap terminals be assessed individually. The possibility
 of genetic testing brings that given the privilege. The other possibility, of course,
 is that the jaunt All this undermines the possibility of economic reform and requires
 get. (Knowing that there is no possibility of attempting coitus takes the who was
 openly cynical about the possibility of achieving socialism 5 so that they can
 perceive the possibility of being citizens engaged in poisoning and fire, facing the
 possibility of their own death just to be hearing yesterday that the possibility of
 using the agency to gather in 1903, and I don't foresee any possibility replacing
 that. The car we a genetic factor at work here, a possibility supported by at least a
 few refused even to entertain the possibility that any of the nations of the has a
 long history, there is the possibility that the recent upsurge in Police are
 investigating the possibility that she was seen a short time any doctors who think
 there is a possibility that they may have been infected are in a store, there is a
 good possibility that you are wearing moisturizer living must be made. The possibility
 that a young adult will be he'd completed his account of the possibility that there
 was a drug-smuggling has been devoted to exploring the possibility that so-called
 ancient peoples
```

Once such a concordance has been created, the co-occurences of other words with the KWIC can be analyzed. This is done by means of a t-score. If we take for example the word *stranger* (comparative adjective and noun), a t-score analysis will provide us with information such as word frequency in the corpus: words such as "no" and "to" are not suprisingly very frequent; a word such as "controversy" much less. It then calculates the occurences of that word together with the KWIC ("joint frequency") to determine if that combination is unusually common, in other words, if the word combination occurs significantly more often than would be expected by its frequency alone. If so, the collocation is considered strong, and is worth paying closer attention to. (For an example of how this is done, see the collins cobuild concordancer and collocation sampler.

In our example, "no stranger to" is a very frequent collocation; so are words such as "mysterious, handsome, and dark". This comes as no surprise. More interesting, however, is "no stranger to controversy". Perhaps the most interesting example, though, is the idiomatic "perfect stranger". Such a word combination could not be predicted on its own, as it does not mean "a stranger who is perfect" as we should expect. Its unusually high frequency shows that the two words collocate strongly and as an expression are highly idiomatic.

The study of corpus linguistics provides us with many insights into the real nature of langauge, as shown above. In essence, the lexicon seems to be built on the premise that language use is best approached as an assembly process, whereby the brain links together ready-made chunks. Intuitively this makes sense: it is a natural short-cut to alleviate the burden of having to "re-invent the wheel" everytime we speak. Additionally, using well-known expressions conveys loads of information rapidly, as the listener does not need to break down an utterance into its constituent parts. In *Words and Rules* Steven Pinker shows this process at work with regular and irregular verbs: the former is collected and provides us with rules which can be applied to unknown words (for example, the –ed ending for past tense verbs allows us to decline the neologism "to google" into "googled"). Other patterns, the irregular verbs, are stored separately as unique items to be memorized.

3. Idiomaticity

Recent research by Wray (2002) and others point to the largely idiomatic nature of language. This implies that language is not created anew each time it is used, but rather based on conventions, patterns, and partially-fixed expressions. Wray argues that language ranges from the purely idiomatic ("pick up the tab" for "pay the bill") to formulaic or semi-fixed expressions ("I look forward to seeing / hearing from / meeting you"). The reason for this is quite simple: the Lexicon works on a "Needs Only Analysis": each new addition to the Lexicon is analyzed for recognizable patterns; if they are found, the new item is stored away under "known" – no further analysis is needed. If, however, the item is alien, a known rule has to be altered to fit the newcomer, or it is stored away as an "idiom": in other words, an item that has an independent meaning of its own. This line of argumentation follows that of Pinker's very closely, only that the difference lies at sentence-level and not individual words.

A good example of NOA at work is shown in studies done on the few native speakers of Esperanto, the famous, artificially created language, which has an entirely regular grammar system on the assumption that this would make it easier to learn for everybody to learn. But something interesting happened as it was being learned by young "native speakers": as they were learning it, they "improved" on the artificially created rules. This interesting development suggests several things:

1. Esperanto was treated as a pigdin (a simplified, watered-down version of an organic language). The first native speakers were turning it into a creole.

2. Language acquisition for native speakers does not work on a top-down principle of clear grammar rules which we all strive to adhere to, but rather bottom-up, learned in bits and pieces, with at best a partial command of all its rules.

3. Language is organic and "needs-only": new items are analyzed ad hoc and adapted into existing structures or stored separately as independent entities.

One example might help to illustrate this point: telling the time. This is often purely idiomatic. For example, we say "a quarter past three" – though not "three-quarters past three". Three-quarters of a pint, on the other hand, is perfectly acceptable. The same applies to the use of 1/3: "a third past three", is not acceptable, whereas "a third of those who voted" is. The use of numbers and fractions in other contexts, however, is highly regulated and follows fixed patterns – though even here numbers show certain frequencies and preferences for other words.

The Lexicon does not pause long to analyze such discrepancies: instead, it analyzes new items for rules; failing to find them, it stores them away pragmatically in a new file, then moves on.

4. Metaphor as an Organizational Principle

Another method of effective language storage in the Lexicon includes the use of metaphor as a storage principle. ("Storage" and "files" are good examples of how human memory and computer memory have been linked to the same vocabulary; this was not always the case). Lakoff's work (1980) is usually cited as the cornerstone to studies of metaphor in the language. One example is quite common: "time is money". We can *save, spend* and *waste* both *time and money*. Another interesting example comes from business and sex: businesses penetrate the market, attract customers, and discuss "relationship management." Business is also war: launch an ad campaign, gain a foothold in the market, suffer losses. Systems, on the other hand, are water: a flood of information, overflowing with people, flow of traffic. The NOA theory of Lexicon acquisition argues that the metaphoric sorting filter helps to simplify language storage and avoid overload.

5. Grammar

Enough work has been done on grammar without needing to discuss it further here. But even here computer research has revealed that grammar – in the sense of its ability to create entirely new language – is avoided as far as possible. Biber and his team working at the University of Arizona on the Cobuild GSWE noted an unusually high frequency of word bundles that, on their own, lack meaning. But a sample of one or two quickly suggests their function: they can be inserted as grammatical glue without any prior analysis of form. Even a cursory observation of examples reveals how commonplace they are in all forms of language use, yet we are hardly aware of their existence. Research suggests that language is heavily peppered with such bundles in all registers; two examples include *do you want me to*, commonly found in speech, or *there was no significant* found in academic registers. Put together in speech, they can create comprehensible sentences, such as *I'm not sure* + *if they're* + *they're going* to form *I'm not sure if they're going*. Such a sentence eases the burden on the Lexicon as it requires no grammatical analysis whatsoever.

6. Register

Michael K. Halliday (1987) proposes a useful dichotomy of spoken and written language which actually entails a shift in paradigm: while linguistic theory posits the superiority of the spoken language over written language (as the former is the origin, comes naturally, and thus precedes the written language), or the written over the spoken (for the same reasons: the written language being the highest form of rudimentary speech), Halliday states they are two entirely different entitities. In short, he claims that speech is grammatically complex while writing is lexcially dense (Halliday, 1993). In other words, a sentence such as "a cousin of mine, the one who I was talking about the other day –the one who lives in Houston, not the one in Dallas – called me up yesterday to tell me the very same story about Mary, who…" is most likely to be found in conversation, not as a newspaper headline. "Prime Minister vows conciliation", on the other hand, would be a typical news headline.

Halliday's work suggests something radically different: language behaves in registers. Biber et al working on the LGSWE worked with four (these are not exhaustive, merely exemplary): conversation, literature, news, academic. These four registers clearly highlight distinctions among language use which would not be clear through a "grammatical" approach. Not surprisingly, each register favors the use of different words and structures: whereas news headline stories, for example, are grammatically simple, conversational anecdotes are full of lexical repetition. The lexis of the news, however, can be quite dense, just as the grammar of speech can be incredibly complicated.

The upshot of this finding lays the theoretical groundwork for a linguistic study of language behavior in particular environments. This work has just begun.