In:New Trends in Modern Linguistics, 8 (Annual catalogue, Maruzen Publishers, Tokyo) (2000).

1999年度スピノザ賞受賞記念講演全文（英訳）

How the ear comes to hear

Anne Cutler
Director of Max-Planck-Institute for Psycholinguistics, Netherlands
Winner of the SPINOZA AWARD 1999

The Word

There are an inordinate number of words. Every language has a stock of tens of thousands, perhaps hundreds of thousands of words. Languages are much more economical with speech sounds (phonemes). English, for example, relies on some 41 different phonemes, and Dutch uses 35, both totals somewhat above the international average. (Japanese, on the other hand, is much more representative, being very close to the world mean!) This means that in every language a great many words are built up out of a very small number of phonemes. As a direct consequence, words are very similar. *Ear, hear, gear, near, rear; house, mouse, louse, douse; take, tape, table, tailor*: these are all English words. Furthermore, short words are often embedded in longer words. Every *rear, gear* or *tear* hides an *ear*. There is an *ape* in every *tape* and *shape*. *Stay* is a word, but it can also become *state, steak*, or *stain*; *stay* can be found in *estate*, or *mistake* or it can continue as *status* - and so forth. This means that in every utterance there is good chance that an unintended word will inadvertently appear in the stream of speech. Sometimes the utterance is actually ambiguous (in Dutch we have *voor mij is er geen luis te raar* - for me no louse is too strange - *voor mij is er geen luisteraar* - for me there is no listener), and sometimes there is only one meaning, despite the embedded word (*voor mij is er geen pluis te raar* - for me no fluff is too strange).

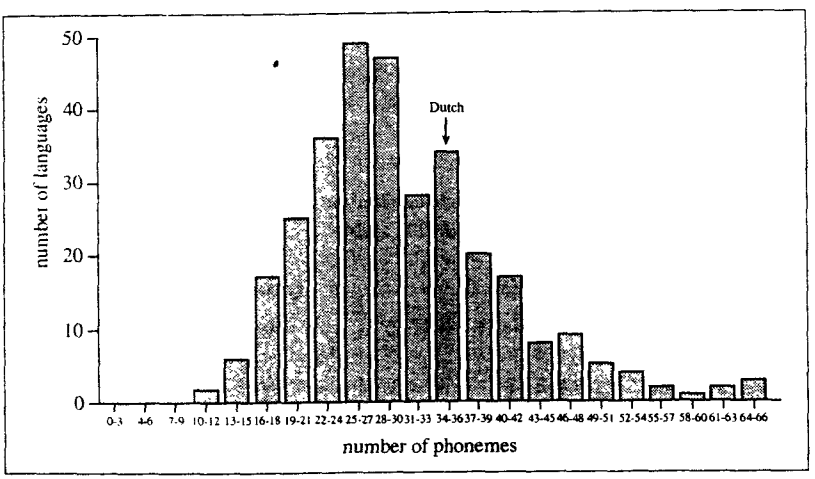

Figure 1.

These coincidences between words and word-segments wouldn't be relevant to the listener if speakers would just give clear signals to indicate where one word ended in their spoken utterances, and the next began. Sadly, speakers don't offer their listeners this particular service. Spoken language is continuous, words blend into one another without interruption and there are hardly any indicators that signal a word boundary. It is up to the listener to recognise the words that the speaker intended, and to exclude the unintended words inadvertently introduced.

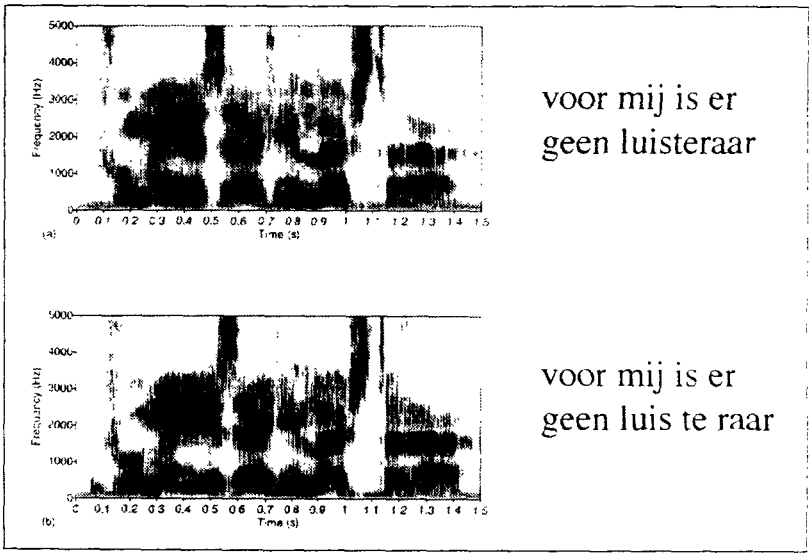

Figure 2

The Ear

The ear of the listener receives a continuous speech-signal, without boundaries between intended words, and including a lot of words that were never intended. In our experience (as average listeners) this becomes a tidy series of separate words, each one neatly following the next. The listener doesn't perceive the continuity of speech, isn't worried by the unintended words, and doesn't notice them at all (except perhaps the occasional inveterate punster).

Investigating how it happens that only the intended words are recognised, i.e. how listeners effortlessly solve an apparently difficult problem, is one of the nicest tasks in psycholinguistics. It goes without saying that this task demands some ingenuity, because it requires, like every other aspect of our whole field, the making visible and measurable of processes which not only occur within our heads but also proceed very rapidly. Sadly, we haven't got a window in the head through which we can observe these rapid processes. So psycholinguists have to resort to indirect methods.

These indirect methods include a variety of simple tasks that we get subjects - average listeners - to carry out in psychological laboratories. In one such task we let the listener hear a series of non-existent words: *thoople, larmage, lunchaf, crinthish.* A number of these non-words conceal real existing words, and the task of the listener is to detect these embedded words and then, as rapidly as possible, to push a button and say the detected word. The listener doesn't know in advance which words will come up (in this respect the situation resembles a normal conversation!), and most of the non-words do not include any kind of word. (In the non-word examples above, we hope that the listener will find *lunch* in *lunchaf* - and nothing else.)

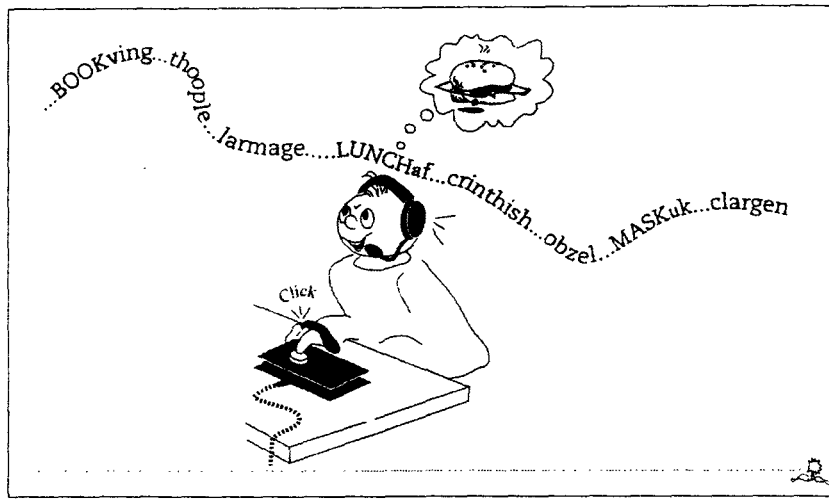

Figure 3

It is clear that these methods give us a glimpse of the process of word recognition in continuous speech: the word presented is embedded in a nonsense context that blends uninterrupted into the real word (just like the neighboring words in normal speech). The measurement of the reaction time - how fast (or slowly!) the listener detects a specific word and signals this detection by pressing the button - offers us a way of comparing the relative difficulty of different contexts. Contexts vary considerably in difficulty. One context can contain more unintended words than another. Or sometimes there are phonemes in the context that are impossible to combine with the word, so that the word springs out at you.

A few examples. Suppose that a listener is presented with the following words: *lunkime bafbege fooprock*··· The reliable test subject finds a known word ie. *rock*, in the third non-word, presses the button, and promptly says "rock". Another listener is given almost the same speech, with a small difference: *lunkime bafbege foomrock* ··· and this second subject reacts noticeably faster than the first.

That is, *rock* is easier to find in *foomrock* than in *fooprock*, and the reason is that the two phonemes [m] and [r] may not be combined in one English word: there is no English word in which a syllable begins or ends with [mr]. So there has to be a boundary between these two phonemes, and this boundary coincides

with the beginning of *rock*. From this type of experiment we know that listeners can make very rapid use of this sort of sequence restriction.

In a similar sort of experiment we have compared *mintayf* and *mintowf*, a comparison of contexts *-ayf* and *-owf* for masking the word *mint*. To be honest neither of these contexts makes it easy for the listener to find the word *mint*; this is because the [t] in *mint* combines happily with the subsequent vowel. But *-ayf* is nevertheless harder than *-owf*. Why? The reason is that a great many English words (*tailor, tape, table* etc.) begin with *tay*, while there are only a few that begin with *tow-* Evidently hearing *ta-* is enough to activate many possible candidate words. This makes it more difficult for the listener to recognise that the [t] actually belongs to *mint* and has nothing to do with the following vowel.

What we see here is a "competition-effect" - a contest between the words that are (wholly or partly) compatible with the speech-stream. *Tailor, tape, table* and the other words vie with *mint* for the single [t]. For the same reason the word *less* is easier to find in *boless* than in *choless* and *ham* easier in *hambup* than *hambur*; in *choless* and *hambur* there is competition with the words *cholesterol* and *hamburger* which makes it more difficult to detect *less* or *ham*. This happens even though the listener in the experiment knows that all the non-words consist of only two syllables, and that all the embedded words are single syllable. Thus the words *cholesterol* and *hamburger* can never appear. Despite this we see faster reactions in *boless* and *hambup* than in *choless* and *hambur*, which can only be explained as effect of undesired competition from *cholesterol* and *hamburger*.

From these and many similar experimental results, a picture of the process of word recognition begins to emerge. Words that appear in the speech signal are activated in our head. The activation process is automatic and can be set in motion by a portion of the word. That is, the unintended words that reach our ears are sometimes actually activated. If we hear *tay* then *tailor, tape*, and *table* are all made available, and *stay* activates *stay, steak, status, state* and so forth. The activated words energetically vie with one another, and this sort of contest can slow the recognition process (by several milliseconds!). The winner of the competition is the word that gets the best support from the speech signal (the signal *state* naturally offers more support to the word *state* than to *sta, steak*, or *status*), or they are the words that together best fit the whole speech signal (*stay together* temporarily supports *state* more than *stay*, but eventually the [t] is won by *together*, so that *state* loses that extra support and *stay* emerges as the winner).

Despite this, we as listeners are not entirely at the mercy of our vocabulary and the battles that take place within it. Happily, we have an armoury of procedures which can almost immediately reject inadvertent words. The operation of such procedures can again be illustrated by the same sorts of experiment.

Suppose that the non-word *prock* is presented. *Prock* conceals *rock*, but here *rock* is hard to find, harder, for example than in *fooprock*. Why? In this case it doesn't depend on whether or not the sounds can combine, because *rock* isn't just difficult to recognise in *prock*, but also in *mrock*, when compared with *foomrock*. This is significant, because, as we saw above, an [m] before an [r] marks a definite boundary. A [p] before an [r] doesn't constitute such a marker; the [p] with the [r] and the [o] can be the beginnings of such words as *prod, prop* and *protestant*. Despite this, the [m] of *mrock* like the [p] of *prock* represents a difficult context. The reason for this must lie in the form of the context: *foom* and *foop* are syllables, while

[m] and [p] are just consonants. Consonants are very useful sounds when it comes to distinguishing one word from another, like *take* from *tape* or *hear* from *near*, but what they can't do is constitute a word. Vowels can - think about *a*, *eye* - but consonants can't. An embedded word is in fact easy to find in a context that contains a vowel, but it is always difficult to find if the context consists of just one consonant.

Clearly, this represents a sort of check: if an activated word leaves the rest of the expression as something that can't be another word, then there is little chance that the activated word indeed constitutes part of the message. This test serves as a simple method to reject unintended (but nonetheless automatically activated) words and thus to minimise undesired competition effects.

Embedded words that leave unviable remainders don't have to stay activated; the listener can throw them out immediately. And to test whether a bit of speech is viable as a word, you only have to ask: does it contain a vowel? If there is a vowel, then it can indeed be a word (to be sure *foom*, *foop*, *cho*, *af* and so forth are not English words, but they could have been English words). Without a vowel there is no viability: [m] and [p] are not only not words, but also they could never be words.

Thus, if a Dutch speaker hears *voor mij is er geen pluis te raar*, then the activation of *luisteraar* can be immediately annulled , because the residual [p] can't be a word. An English speaker can reject *metaphor* in *met a fourth time* just as quickly because the residual [th] from *fourth* can't be a word. And when we hear *hear* we don't have to pay serious attention to the activation of *ear*, because *ear* leaves [h], and [h] can never be a word. The *ear* does its best, but *hear* is what is heard.

Acknowledgement

This is the translated text of a speech "Hoe het woord het oor verovert" delivered in the Nieuwe Kerk in The Hague on February 15th, 2000, on the occasion of the presentation of the Spinoza Awards 1999. The English translation of the Dutch original text is by A.W. Sloman.