Shannon vs. Chomsky: Brain Potentials and the Syntax-Semantics Distinction

Mathias Winther Madsen

Institute for Logic, Language, and Computation

Abstract. The N400 and the P600 are two specific patterns in the electrical brain potential which can sometimes be found when people read unexpected words. They have been widely claimed to be the neurological correlates of semantic and syntactic anomalies, respectively. Evidence from the last decade has raised some serious doubts about that interpretation, however, and in this paper, I first review some of this evidence and then present an alternative way to think about the issue. My suggestion is built directly on Shannon's concept of noisy-channel decoding by tables of typical sets, and it thus is essentially statistical in nature. I show that a proper application of Shannon's concepts to the reading process provides an interesting reinterpretation of our notion of "syntax," thus questioning some fundamental assumptions of linguistics.

One of the key ideas that helped establish linguistics as a respectable science in the 1950s was the idea that the processing of syntax was an autonomous mental module that could work independently of the processing of meaning. Noam Chomsky famously pushed this point by exhibiting sentence pairs that seemed to dissociate the two kinds of violation [Chomsky, 1957, 15–16]:

- The child seems sleeping (syntactic anomaly)
- Today I saw a fragile whale (semantic anomaly)

Based on such examples, he concluded that violations of syntactic rules had nothing to do with the specific meaning of the word in the sentence, and thus

> we are forced to conclude that grammar is autonomous and and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure. [Chomsky, 1957, 17]

While the introspective case from looking at example sentence was Chomsky's strongest argument at the time, he did imagine hypothetical psychological experiments that could be used to corroborate his claim, including reading or memory tests that could distinguish "ungrammatical" sentences from mere nonsense [Chomsky, 1957, 16].

This paper is about one attempt to realize such a research program. As I shall explain shortly, the brain sciences found some new results in the 1990s which seemed to finally put Chomsky's claims on a solid scientific basis. For a

period of about ten years, the consensus was accordingly that the anatomical and psychological reality of Chomsky's little syntactic engine was a scientific fact. Syntax had acquired the authority of being "brain stuff."

In one sense, this paper is an elaborate criticism of this idea. I will review a wealth of empirical evidence from the last decade which seriously complicates the picture of the brain phenomena that were originally used to vindicate the Chomskyan view of language. Even a casual look at this literature exposes an abundance of unruly little details that fit very badly into the old pigeonholes.

But this is also a constructive paper: I propose a precise computational model that may explain most of the psychological phenomena that I will be dealing with here. This model is based on a seemingly insignificant idea about statistical error used by Claude Shannon in the 1948 paper which founded the field of information theory [Shannon, 1948].

This idea has some deep connections to the foundations of statistical inference and thus to the notion of rational choice and behavior in the most general sense. Presenting my model will thus require a detour around some basic concepts from information theory and statistics. Once I have introduced those concepts, I can focus on a more specific example of a decoding problem very similar to the experimental situations used in the psychological literature. I conclude the paper by a brief discussion of the merits and flaw of this proposed model.

1 The N400 and the P600

In this section, I will introduce the brain phenomena that are pivot of the entire paper. These phenomena are a couple of distinctive changes in the electrical brain potential that can be measured off the scalp of a person's head while he or she is reading. I will discuss the difficult issue of the interpretation of these modulations of the brain potential in more detail in the next section.

1.1 The N400

Suppose I present you with a sentence by flashing each word on a computer screen at regular intervals:

– I . . . spread . . . the . . . warm . . . toast . . . with . . .

While you are reading the sentence, I read record the electrical activity at different places on the top of your head. Since all nerves in your body change the electrical field around them when they are excited, such a recording provides an estimate of how active the nerves inside you head are. Assuming that electrical activity in the head means mental activity, an experimental paradigm like this can thus provide a rough indication of which words in a sentence that require a lot of cognitive effort, and which that don't.

This means that rather specific differences in the form and content of sentences can be compared. I could, for instance, change a single word in a sentence and then compare the pattern of electrical activity produced by the original sentence and the manipulated one:

– I spread the warm toast with butter (baseline)
– I spread the warm toast with socks (manipulation)

This manipulation was the idea behind an experiment which was first carried out by Martha Kutas and Steven A. Hillyard in 1980, and since then replicated many times. Their experiment involved a set of sentence pairs like the ones above, differing as to whether final word was normal and expected, or "semantically inappropriate (but syntactically correct)" [Kutas and Hillyard, 1980, 203].

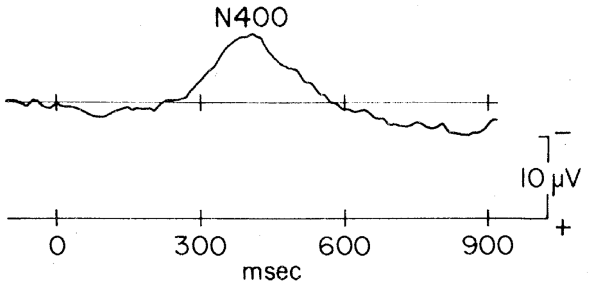


Fig. 1. A graph from Kutas and Hillyard (1980: 204) showing the averaged response to the inappropriate word minus the averaged response to the appropriate. Both waveforms were recorded by an electrode at the back of the head (the medial parietal position). As is conventional in electrical engineering, negative charge is plotted as up.

By averaging over a lot of subjects' responses to the two versions of the sentences, Kutas and Hillyard found that the oddness of the manipulated sentences tended to provoke quite specific electrical response characterized by an excess of negative electrical potential compared to the baseline condition. Because this bump on the graph showed up around 400 milliseconds after the unexpected word was presented, they called this component of the waveform "N400." They suggested that it might be a kind of "second look" effect, a cognitive equivalent of rubbing your eyes in disbelief.

1.2 The P600

Kutas and Hillyard's experiment was explicitly designed to manipulate "semantic congruity" while keeping syntactic factors constant. This naturally led to the question of what effect the opposite manipulation would have.

This question was answered in 1992 by Lee Osterhout and Phillip J. Holcomb, who compared a different kind of sentence pairs, but otherwise used a virtually identical experimental paradigm:

– The swimmer decided to lose weight (baseline)
– The swimmer urged to lose weight (manipulation)

Although the difference in wording between these two sentences lies in the verb (*decided* vs. *urged*), the difference in how odd they are is only apparent later: The sentence fragment *The swimmer urged ...* can be completed in many perfectly fine ways, but the extension *The swimmer urged to ...* cannot. It is thus only when the word *to* is presented that it starts to get difficult to come up with a plausible structural interpretation of the sentence.

Consistent with this reasoning, Osterhout and Holcomb found a marked difference between the reaction to the word *to* in the two conditions, with the manipulated sentence provoking an increased positive potential. They found that this positive component of the response peaked at about 600 milliseconds after the onset of the word and logically named it "P600."

Having thus apparently found a neurological correlate of the mismatch between verb type and complement type, Osterhout and Holcomb speculated that "the P600 and N400 effects are elicited as a function of anomaly type (syntactic and semantic, respectively)," and that "the P600 co-occurs with syntactic anomaly" [Osterhout and Holcomb, 1993, 785 and 798].

Their results thus seemed to support the idea that syntax and semantics are distinct features of language — not only in the mind of the linguist, but also in the brain of the language user.

2 The Frayed Ends of Syntax

With the discovery of the P600, the brain sciences produced a strong neurological argument in favor of the Chomskyan view of language. Linguists could now with confidence postulate the "autonomy of syntax" and cite the authority of brain science as support. The issues raised by Chomsky from a purely introspective perspective seemed to be finally empirically settled.

The issue is, however, not quite as simple as this. Within the last decade, a number of experiments have documented P600 responses to sentences that are not syntactically anomalous in any usual sense of the word "syntax." I will spend the remainder of this section illustrating these complications in some detail, drawing quite heavily on an excellent review paper by Gina Kuperberg [2007] in the process.

2.1 Gaps and Plugs

One of the first indications that the P600 could not unproblematically be equated with broken syntax came from a Dutch study which manipulated the test sentences by changing an auxiliary verb from *was* to *has* [Hoeks et al., 2004].

In Dutch, the main content verb is often placed at the very end of the sentence. If you read such a verb-final sentence in a word-for-word presentation, a whole string of nouns and auxiliary verbs can thus be presented to you before you finally reach the main content verb.

This means that one can construct a Dutch sentence such that it creates a very strong expectation as to what the final verb will be, and those expectations can then either be respected or violated. In the materials used by Hoeks et al., a number of sentences were thus manipulated either by substituting the main verb, or, as in the following pair, an auxiliary verb:

- De eed werd door de jonge artsen afgelegd. ("The oath was taken by the young doctors.")
- De eed heeft de jonge artsen afgelegd. ("The oath has taken the young doctors.")

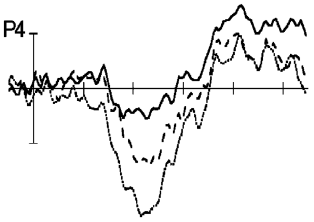


Fig. 2. A single-electrode recording from Hoeks et al. [2004, 68], showing the manipulated conditions minus the baseline. The strongly positive graph corresponds to the sentence with *has* instead of *was*.

The manipulations of the main verb (e.g., substituting *save* for *take*) tended to produce the expected N400 effect. The substitution of *has* for *was*, however, triggered an extreme excess of late positive potential, beginning around 600 millisecond after the final verb was presented (cf. Fig. 2).

Of the two strange but grammatical sentences, one would thus produce an N400, and the other a very pronounced P600. This ran directly counter to the notion that the P600 exclusively tracks syntactic problems.

Similar effects were reported for English. In one 2003 experiment, Kuperberg et al. compared two different kinds of unexpected verbs and again found that only some of them would provoke a N400 response:

- For breakfast, the boys would only eat ... (baseline)
- For breakfast, the boys would only bury ... (N400)
- For breakfast, the eggs would only eat ... (P600)

Kuperberg et al suggested that this difference might be explained in terms of the thematic roles prescribed by a verb like *eat*, assuming that "the P600 is sensitive to violations in this thematic structure" [Kuperberg et al., 2003, 127].

This line of thought was also supported by another study which compared bizarre subject-verb combinations to mere nonsense [Kim and Osterhout, 2005]. This study, too, found marked differences in the kind of response elicited by the two manipulations, compared to a baseline of ordinary English:

- The library books had been borrowed by the graduate student. (baseline)
- The tragic mistake was borrowing in a hurry. (N400)
- The library books had been borrowing the graduate student. (P600)

Like Kuperberg et al., Kim and Osterhout suspected that the difference between the two target sentences had something to do with how nouns plug into verbs. They formulated this intuition in terms of a "semantic attraction to particular predicate–argument combinations" [Kim and Osterhout, 2005, 215].

2.2 Where Does Grammar End?

Looking at these examples, one might come to expect that the crucial issue here had something to do with animacy: *boys* are animate and therefore appropriate subjects for the verb *eat*, but *eggs* are not.

However, as several authors have argued, this distinction does not quite capture the differences in the empirical material [Kuperberg et al., 2003, Kuperberg, 2007]. Both *library books* and *tragic mistakes*, for instance, are inanimate nouns and thus inappropriate as subjects for *borrow*. Still, the former choice elicits a P600, while the second an N400, as mentioned above.

It should also be kept in mind that many earlier experiments had used subject-verb combinations that were clearly illegitimate in terms of animacy or other selection criteria, but still produced an N400 and no P600. One study for instance used the following target sentences:

- Der Honig wurde ermordet.
  ("The honey is being murdered.")
- Der Ball hat geträumt.
  ("The ball has dreamt.")

Both of these sentence types provoked a marked N400 response, but no P600 [Rösler et al., 1993]. It would thus appear that inanimate objects can quite easily be used with verbs that require animate subjects without provoking a P600 (although it will very likely produce an N400).

As for the opposite direction, other research groups have also documented P600 responses to sentences with animate subjects doing animate things. One study by Kolk et al. [2003], for instance, found this in Dutch sentences like

- De vos die op de stropers joeg ...
  ("The fox that was hunting the poachers ... ")
- De kat die voor de muizen vluchtte ...
  ("The cat that was fleeing the mice ... ")

In a later experiment applying essentially the same materials, van Herten et al. [2005, 249], to their own surprise, also found P600 effects for sentences such as

- De bomen die in het park speelden ...
  ("The trees that played in the park ... ")

In a later paper, they hypothesized that a predictor of this P600 effect could be that the verb and noun which are stuck together tend to to occur in the same contexts, but not in subject-predicate combinations [van Herten et al., 2006]. This would explain difference between sentences such as the following two:

- Jan zag dat de olifanten de bomen snoeiden en ...
  ("John saw that the elephants pruned the trees and ... " — P600)
- Jan zag dat de olifanten de bomen verwenden en ...
  ("John saw that the elephants spoiled the trees and ... " — N400)

There are thus quite strong reasons to doubt that the conditions producing the P600 is a matter of plugging an inanimate noun into a frame that requires an animate, or vice versa. A wide variety of relationships between words seem to be responsible for turning "semantic" anomalies into potentially "syntactic" ones.

2.3 Grammatically Ungrammatical

Given all of this data, it seems that there is no clear and straightforward relationship between the P600 and the arguments that can be plugged into a verb. The issue does seem to be syntactic well-formedness, nor does it seem to be about animacy or thematic structure. It might still be the case, however, that some other grammatical parameter — say, aspect, mood, gender, number — is the key. But even this vague suggestion has some empirical evidence against it.

One such piece of evidence comes from a study that recycled some sentences which had already previously been shown to produce strong and reliable N400 effects. In the new experiment, an elaborate context was prepended to these sentences, and the brain potentials were recorded after the same word as in the previous study [Nieuwland and Van Berkum, 2005]. One of stories used in this experiment read as follows, in translation from the original Dutch:

– A tourist wanted to bring his huge suitcase onto the airplane. However, because the suitcase was so heavy, the woman behind the check-in counter decided to charge the tourist extra. In response, the tourist opened his suitcase and threw some stuff out. So now, the suitcase of the resourceful tourist weighed less than the maximum twenty kilos. Next, the women told the suitcase that she thought he looked really trendy. ...

Surprisingly, this long build-up to the crucial unexpected word completely canceled N400 effect and instead produced a strong P600. The manipulation of the context which made a suitcase a natural prop thus dramatically changed the way the subjects read the nonsense sentence *the woman told the suitcase*.

But perhaps the deepest problem with the idea that the P600 can be described in grammatical terms is that it can be provoked by sentences without any syntactic anomalies at all. Paradoxically, this was already shown in a study from 1999 [Weckerly and Kutas, 1999] which compared sentences of the following form:

– The novelist that the movie inspired praised the director for ...
– The editor that the poetry depressed recognized the publisher of the ...

At the time, this Weckerly and Kutas used these findings in order to make a point about "the use of animacy information in the processing of a very difficult syntactic structure" [Weckerly and Kutas, 1999, 569]. There are several other plausible ways to explain their data, though. It could be, for instance, that Weckerly and Kutas' subjects were simply so used to novelists that inspire movies and editors that edit poetry that when the action started flowing in the other direction, the reading process had a hiccup.

Regardless of these points of interpretation, however, it seems clear that there are some serious problems with seeing the P600 as the exhaust pipe of a brain's syntactical processor. There are several kinds of semantic oddness that are quite strongly related to the P600, and as we have seen, it can be quite heavily modulated by discursive context and even be elicited by grammatical sentences. The standard tools from the toolbox of theoretical linguistics thus seem to have some problems getting to grips with the phenomenon at hand.

I take this as a sign that we need to back up a bit and take a fresh look at our assumptions. In the next section, I will consequently suggest that in order to understand the brain processes behind the N400 and the P600, we need to go all the way back to Chomsky, and then go a little further back.

3 The Statistics of Decoding: Errors of Type I and II

In the preceding section, I have presented a number of empirical phenomena that seriously challenges the traditional identification of the N400 and the P600 with semantic and syntactic anomaly, respectively. As the examples have shown, the P600 in particular crops up in a number of different circumstances that cannot be equated with broken syntax with stretching our notion of syntax by any stretch of the imagination. This raises two questions.

First, we might wonder whether there any system at all to where the P600 is present. Indeed, after looking at the wide variety of examples in the previous section, one might get the impression that brain potential just aren't the kind of phenomena that can be adequately predicted. It is an open question whether we can even describe the phenomenon in intuitive terms, or whether we can look at a sentence and guess what kind of brain response it will evoke.

Second, assuming that there is some kind of system, the question is what formalism that would be most suited to articulate it. It is tempting to pull down the grammatical vocabulary from the shelf because, after all, we are talking about language; but it is not a given that a theory of the N400 and the P600 should come in such a form or correspond in any way to Chomskyan linguistics.

In this section, I will suggest that there is indeed a way of understanding what the two brain components of the reading-related brain potentials are, but they have little to do with language as such. Instead, I will draw on some classical statistical insights from information theory [Shannon, 1948] and approach the reading situation as a kind of decoding problem. The two brain responses will then show up as correlates as particular kinds of decoding error.

This line of thought is not completely alien to the researchers working in the field. Several authors have had an intuition that the P600 had something to do with "recombination" [Kuperberg, 2007], and that the difference between the N400 and the P600 thus was a matter of whether the sentence seemed worth repairing. Kolk et al. [2003], for instance, noted that both responses were triggered by strange and unexpected events, and commented:

> The problem with such events is that they can have two sources. They can be real, in the sense that an unexpected event has indeed occurred.

> On the other hand, they can also stem from a processing error. [Kolk et al., 2003, 31]

To paraphrase this in information-theoretic terms, strange things can either happen in the source or in the channel. I agree completely with this intuition, and the purpose of this section is to spell it out in some more mathematical detail. My goal is thus to explicate some theoretical ideas and assumptions which have mostly been implicit in the psychological literature (presumably because of the brevity and specificity constraints in that tradition).

The concept of statistical error that I present here is based on an idea used by Shannon in the proof of the discrete version of his channel coding theorem [Shannon, 1948, Th. 11]. It was later made more explicit by other authors, and I rely in particular on Cover (1975; see also Cover and Thomas, 1991, ch. 7.7). Before I go into this main topic, however, I will first have to discuss a few other basic concepts of information theory.

3.1 Typical Sets and Most Probable Sets

One of the key concepts that Shannon defined was that of the "typical set" [Shannon, 1948, Th. 3]. The typical set of an experiment is the set of outcomes whose logarithmic probability is close to the average logarithmic probability.

To put this differently, suppose we identify the experiment in question with a random variable X and use label $p(x)$ to mean the probability of the event $X = x$. As suggested by Samson [1951], we can then think of the number $-\log p(x)$ as a quantification of how "surprising" the observation $X = x$ is. This surprisal value is then a number between 0 and ∞, with less probable events having a higher surprisal.

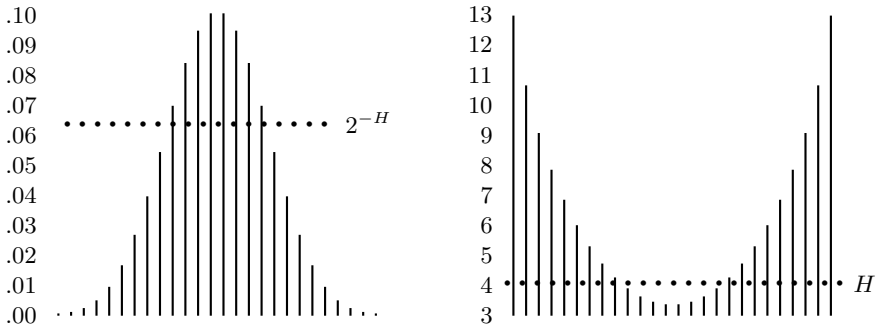


Fig. 3. Left, the distribution of the sum of 5 dice throws, showing the typical probability 2^{-H} as a dotted line. Right, the same distribution plotted on an inverse logarithmic scale.

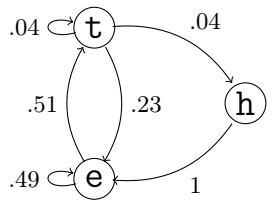
A random variable X that can take different values with different probabilities is thus always associated with an abstract "surprisal" variable $-\log p(X)$

which ranges over the set of surprisal values. The expected value of this surprisal variable is $H = H(X)$, the entropy of X (cf. Fig. 3).

Using these concepts, we can then reformulate the definition of the typical set as the set of events $X = x$ whose surprisal value is close to the average surprisal. For a fixed $\varepsilon > 0$, the ε-typical set associated with a random variable X thus contains the x for which

$$\left| \log \frac{1}{p(x)} - H \right| \leq \varepsilon.$$

The ε-typical set associated with an experiment does not necessarily include the most probable outcome of the experiment. However, as mentioned by Shannon [1948, Th. 3] and explained in more detail by Cover and Thomas [1991, ch. 3.3], the set of outcomes that are less surprising than H usually differs very little from the typical set: If the values of the random variable X are long sequences from an ergodic random process (such as strings of letters or words; cf. Fig. 4), then including in the most probable outcomes will only change the size and total probability of the set slightly. In many important respects, it thus makes little difference whether we define the typical set by the bound $-\log p(x) \leq H + \varepsilon$ or the symmetric condition $H - \varepsilon \leq -\log p(x) \leq H + \varepsilon$.



x	$-\frac{1}{N}\log p(x)$
`ttheeeteeeeeethеeeeeeeteth`	0.71
`ethetheetheetttheeethethеth`	0.74
`etetheetheeeteeeethethethe`	0.49
`theetheethethethеeeethethe`	0.71
`thethеetheetetheteeeeethee`	0.81

Fig. 4. Sequences of length $N = 25$ from an ergodic source. The entropy rate of the source is $H = .80$ bits per symbol.

3.2 Decoding by Jointly Typical Sets

The conditions discussed above define the ε-typical set of a single random variable X. In the context of information transmission, however, we are rarely interested in a single, isolated variable, but much more often in the relationship between two variables X and Y that model a cause and an effect; or more specifically, a transmitted message and a received noisy signal.

In order to study such relationships, it is useful to consider the typical sets associated with the joint stochastic variable $X \times Y$. This set consists of pairs (x, y) that are typical with respect to the joint probability distribution on $X \times Y$, and it is accordingly called the jointly typical set.

In an information transmission context, a jointly typical pair (x, y) thus consists of a high-probability message x and a noisy signal y with high probability

given the message. The set of all such pairs collectively describe the properties of the communication channel modeled by $X \times Y$. A model of written English, for instance, might include (*forty*, *fourty*) as a typical pair, since *forty* is a common word, and *fourty* a common misspelling of it. Other channels will have other characteristic types of confusability (cf. Fig. 5).

This way of looking at the issue suggests that a table of the typical pairs could be used for decoding: When you receive a message y, you just skim through the table, find a pair of the form (x, y), and then return x as your decoded message. Although this hypothetical algorithm is in usually not feasible in the form sketched here, I will continue assuming its unproblematic use in the following because its complexity problems are largely irrelevant to the point I want to make here.

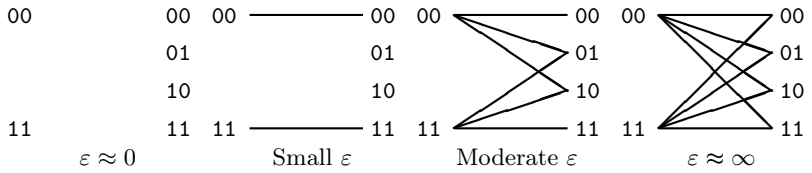


Fig. 5. One possible structure of a jointly typical set for increasing tolerance thresholds and thus decreasing demands on the joint probability $p(x, y)$.

3.3 Two Types of Decoding Error

As is apparent from the description of the hypothetical brute-force algorithm that does decoding by scanning the jointly typical set, two distinct problems can give rise to a wrong answer:

1. There is no x for which (x, y) is typical
2. There is more than one x for which (x, y) is typical

As an illustration of these two types of error, suppose that the message X is one of two points in the plane, A or C, and that the received signal Y is X plus some Gaussian noise (cf. Fig 6). The ε-typical set for this communication channel will then contain the pairs (A, C) for every C on a disc centered on A, and the pairs (B, C) for every C on a disc centered on B. The sizes of the discs are determined by ε, the variance of the noise, and the prior probabilities of A and B.

In a situation like this, the two types of error correspond to different regions of the plane: Errors of the first type correspond to everything outside the two discs, and errors of the second type to their overlapping parts. Outside the discs, it might be the case that one of the events $X = A$ and $X = B$ has a higher likelihood than the other, but both likelihoods will be extremely low. In the overlap between the two discs, on the other hand, the two explanations have

roughly equal likelihood. We could of course reduce the risk of an error of the first kind by choosing a bigger ε (i.e., making the discs larger), but this would also expand the overlap between the discs thus increase the probability of the other kind of error. There is thus a trade-off between accuracy and coverage.

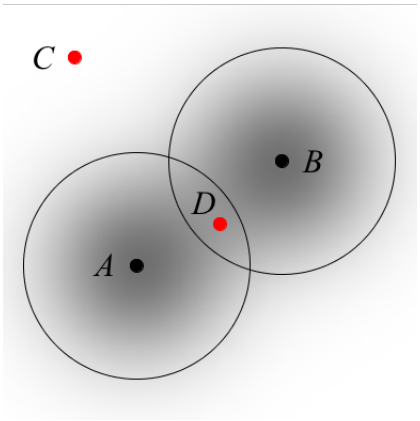


Fig. 6. A inference situation with two possible causes (A and B), a type I error (C), and a type II error (D).

While it is important to notice the similarities and difference between this picture and the frequentist notion of type II and type II errors [Neyman and Pearson, 1928], two differences are worth keeping in mind: First, the Shannon analysis accepts a cause x as an explanation for the observed effect y when

$$p(x \mid y)\, p(x) \geq 2^{-H(X \times Y) - \varepsilon},$$

so the region of acceptance changes with the prior probabilities and the noisiness of the channel. The Neyman-Pearson analysis, by contrast, uses a fixed significance level α and accepts an explanation x when

$$p(y \mid x) \geq 1 - \alpha$$

regardless of the nature of the inference problem. The second point relates more to the philosophies of the two approaches: The purpose of the Shannon analysis is ultimately to assign exactly one cause to each effect, while the purpose of the Neyman-Pearson analysis is to delineate a set of effects that are sufficiently explained by a given cause, disregarding competing alternatives. The two approaches are thus in a certain sense answers to different questions.

4 An Application to Locally Noisy Strings

While the previous section considered some rather abstract statistical concepts, I would now like to turn to a more concrete example which is highly relevant to the general topic of this paper. The example is nominally a statistical model of misspellings, but it applies to any kind of communication channel that can distort a sequence through local changes like reversing the order of two symbols.

Such models have been studied heavily in computer science, and many textbooks contain discussions of similar models. The theory that I employ here thus has roots in classical ideas from computer science, including the notion of dynamic programming [Bellman, 1952, Damerau, 1964, Viterbi, 1967].

4.1 A Generative Story for Misspellings

There are many complicated reasons why people misspell words, but in order to get a better mathematical grip on the situation, I will present a very schematic

model of the process here. The model takes the form of a cartoonish generative story which depicts people like a certain kind of machines that experience some random glitches at crucial points in the writing process. This causes them to fail to correctly externalize the contents of their internal memory. Anyone reading their text will thus have to indirectly infer what their internal memory actually contained, based on the corrupted output signal.

More specifically, let us assume that a writer chooses a word x with a probability proportional to the frequency of the word in written English. Next, the writer consumes the word, letter for letter, in a left-to-right manner (cf. Fig. 7). While traversing this input string, the writer also writes an output string y which more or less faithfully reflects the input. How exactly the writer traverses the word x, and how he or she chooses which letters to output, is decided in a stochastically as summarized in Table 1. Note in particular that the channel model allows for local permutations such as *perc*$\underline{ie}$*ive* instead of *perc*$\underline{ei}$*ve*.

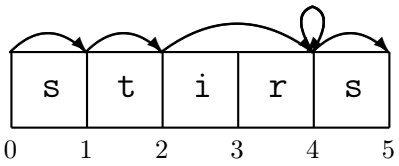


Fig. 7. Consuming the word *stirs* by walking through positions 0, 1, 2, 4, 4, and 5.

State	Action	Probability	Output	Next state
Start	Choose x	$\Pr(x)$	–	0
$i < \|x\|$	Echo	α	x_i	$i+1$
$i < \|x\|$	Change	$\beta/25$	$c \neq x_i$	$i+1$
$i < \|x\|$	Insert	$\gamma/26$	c	i
$i < \|x\|$	Delete	δ	–	$i+1$
$i < \|x\|$	Reverse	η	$x_{i+1}x_i$	$i+2$
$i = \|x\|$	Echo	α	a_i	$i+1$
$i = \|x\|$	Change	$\beta/25Z$	$c \neq x_i$	$i+1$
$i = \|x\|$	Insert	$\gamma/26Z$	c	i
$i = \|x\|$	Delete	δ	–	$i+1$
$i = \|x\|+1$	Insert	$\gamma/26$	c	i
$i = \|x\|+1$	Halt	$1-\gamma$	–	–

Table 1. Generative model for a spelling channel. The rows labeled "Change" and "Insert" should be read as abbreviated forms of several distinct entries, one per possible output letter. Z is the normalizing constant $(\beta+\gamma)/(\beta+\gamma+\eta)$.

As the table shows, the model contains several hyperparameters that can set to arbitrary values, depending on what we think best reflects actual behavior. In the examples discussed below, the values of the hyperparameters were $\alpha = .96$

and $\beta = \gamma = \delta = \eta = .01$, meaning that an average of 24 out of 25 letters are expected to pass faithfully through the spelling channel.

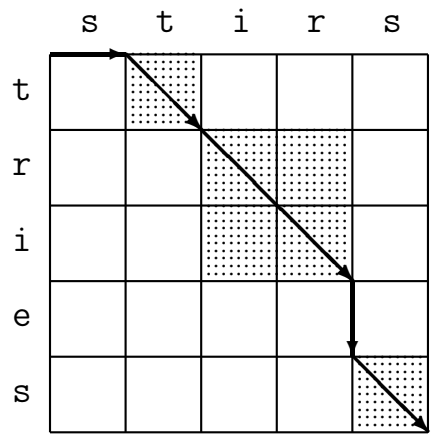


Fig. 8. A path through (*stirs*, *tries*) containing a deletion, an echo, a reversal, a spurious insertion, and another echo.

Thanks to classic computer science trick, the transmission likelihoods $p(y \mid x)$ for this channel can be computed quite easily: Each way of transforming a string x into a string y can be identified with a path through a matrix, with each step of the path corresponding to a particular editing operation like deletion or reversal (cf. Fig. 8). Finding the likelihood $p(y \mid x)$ is thus the same as summing up the probabilities of all paths through this matrix. However, because the transmission errors only have local effects, we can start by finding these sums for all the 2×2 submatrices, then use these results to find the sums for the slightly larger submatrices, and so on. When we reach the size of the original matrix, we have summed up paths.

This style of a bottom-up recursion is what is know as dynamic programming [Bellman, 1952]. I will tacitly use it in all the examples discussed below.

4.2 Estimating Joint Surprisal Values

Analyzing the joint entropy of the channel $X \times Y$ is in fact not completely trivial in the model used here, and the details are not terribly important for my present purposes. It will, however, be very illustrative to see a few representative examples, so I will now provide a rough back-of-the-envelope calculation of some descriptive statistics associated with the spelling channel.

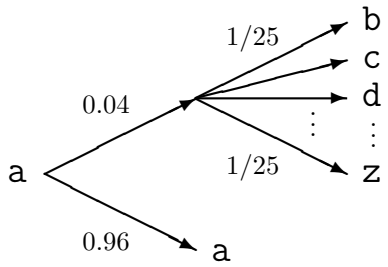


Fig. 9. The simplified channel model.

Let us first assume that an input word x is given and have a closer look at our uncertainty about what the output string y will be. We can think roughly of the writing of a single letter as a combination of two choices: Deciding whether to make a mistake, and if so, deciding which mistake to make. The choice of making or not making a mistake involves two options with probabilities 0.96 and 0.04, re-

spectively. It consequently involves

$$0.96 \log \frac{1}{0.96} + 0.04 \log \frac{1}{0.04} = 0.17 \text{ bits of uncertainty.}$$

In the four percent of the cases where an error is introduced, we further select one of 25 letters, a choice involving an additional $\log_2 25 = 4.64$ bits. However, since this only happens in four percent of the cases, the grand total becomes

$$H(Y|X) = 0.17 + 0.04 \cdot 4.64 = 0.36 \text{ bits of entropy per letter.}$$

An input word with N letters is thus associated with about $2^{0.36N}$ typical output strings, according to this simplified calculation.

The source entropy $H(X)$ only involves choosing a word from the dictionary. Using the frequencies in the Brown corpus [Francis and Kucera, 1967] as estimates of word probability, we arrive at an entropy of about $H(X) = 10.54$ bits. Words like *know*, *while*, *last*, *us*, *might*, *great*, and *old*, have surprisal values close to this average. The most frequent word, *the*, has a surprisal value of 3.84.

To show more concretely what these numbers mean, consider the word *great*. This word has a surprisal value of 10.56 and five letters; we can thus expect an conditional surprisal $H(Y|X = \textit{great})$ in the ballpark of $5 \cdot 0.36 = 1.80$ bits, and thus an average surprisal of about $10.56+1.80 = 12.36$ bits for a pair of the form (*great*, y).

x	y	$-\log p(x, y)$
great	*great*	10.84
great	*graet*	17.32
great	*grate*	24.00
great	*grxqz*	30.42

Table 2. Examples of joint surprisals.

Decomposing this average a little more, suppose the transmission does not introduce any errors; the surprisal at the pair $(x, y) =$ (*great*, *great*) will then be slightly lower than $10.56 - \log 0.96^5 = 10.85$, since this particular output could be produced a faithful reproduction of x, or by a few other much less likely combinations of errors. If a single error is introduced, on the other hand, the surprisal about (x, y) will be about $10.56 - \log 0.96^4 - \log 0.01 = 17.44$ bits, depending a bit on what the error is. Table 2 gives some direct computations that corroborate these estimates.

4.3 Letter-for-Letter Surprisal Values For Competing Hypotheses

All of the preceding discussion assume that an entire word is presented at once. How does the situation change if we suppose that the sequence is revealed letter by letter?

The probability of observing two events, $p(x_1, x_2)$, is the same as the probability of observing the first event and then the second, $p(x_1)\, p(x_2 \mid x_1)$. Since the logarithm turns products into sums, this means that if you observe a string of events, the sum of your individual surprisals are equal to the bulk surprisal you

would experience from observing the whole sequence at once. For instance,

$$\log \frac{1}{p(x_1, x_2, x_3)} \;=\; \log \frac{1}{p(x_3 \,|\, x_1, x_2)} \;+\; \log \frac{1}{p(x_2 \,|\, x_1)} \;+\; \log \frac{1}{p(x_1)}.$$

As an illustration of what this can mean in the context of the spelling channel, suppose that you have forecasted, rather arbitrarily, that you are about to receive the message $x = \textit{fall}$. In fact, however, the actual output that you will see is $y = \textit{flat}$, so at some point along the way, your expectations will be violated. The total surprisal of the pair (*fall*, *flat*) is 25.05 bits, and because of the summation property mentioned above, these 25.05 bits of surprisal must accrue from one or the other letter.

The surprisal need not be evenly distributed, though. When you have only seen the letter *f*, for instance, the hypothesis $x = \textit{fall}$ is still perfectly consistent with the data you have seen. It is only when the next letter turns out to be an *l* that your surprisal jumps upwards. More examples are given Table 3.

x	–	*f*	*fl*	*fla*	*flat*
for	6.73	6.77	13.38	19.03	25.05
large	11.44	18.04	18.11	18.17	24.78
fall	12.73	12.77	19.39	19.39	25.05
flat	13.85	13.89	13.95	14.01	14.07

Table 3. Letter-for-letter surprisals at the string *flat* from the perspective of various hypotheses.

As can be seen from the last row of the table, the hypothesis *flat* terminates at a moderate 14.07 bits, which is not too far from the roughly 12 bits that would expected for a word of this length. For an ε between roughly 2 and 10, this hypothesis would thus be accepted, while the others would be rejected.

Notice also that the table contains occasional cases of reanalysis. In particular, when the *a* in *flat* is revealed, the hypothesis $x = \textit{fall}$ hardly changes surprisal level. This is because the segment *fla* is consistent with a reversal hypothesis under which the *l* and the *a* were written in the reverse order. Observing the *a* thus in a sense explains away the unexpected *l* that would otherwise have to be explained by a spurious insertion.

4.4 Decoding Error in the Spelling Channel

These observations bring me back to my main point, that of the two types of decoding error in a noisy channel. As explained in section 3.3, decoding by typical sets can either break when you have too many or too few candidates that can explain a received signal.

In the context of the spelling channel, an error of type I looks pretty much as we would expect it too: Unrepairable nonsense words like *srxzk* do not look like any known words, and any hypothesis about the underlying message will thus just keep accumulating surprisal as more letters are revealed (cf. Fig. 11). The most probable hypothesis, *size*, eventually accumulates 31.78 bits of surprisal, well above the level expected for a word of this length. The string *srxzk* could

thus lead to a decoding error of type I; a decoder might simply give up making sense of it.

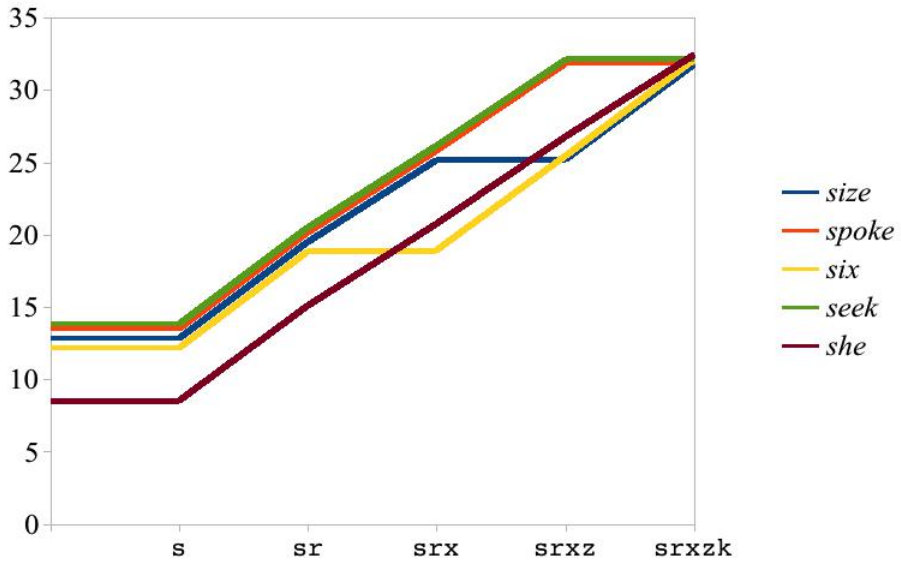


Fig. 10. Letter-for-letter surprisal values for various decoding hypotheses about the two strings *srxzk*.

This situation should be compared to that of the non-word *flide*. As can be seen from the graph, there is a whole clutter of hypotheses which are roughly equally far away from this word. The hypothesis *slide* is the best candidate, but the difference up to the second best, *flies*, is only 0.70 bits. For many choices of ε, either both hypotheses would be rejected, or both would be accepted — both of these outcomes being an error.

From the present perspective, it is also interesting that the graph for the various hypotheses cross each other so frequently: After the fragment *fli* is presented, the two hypotheses *slide* and *fled* are practically equiprobable. After the fragment *flid* is presented, the hypothesis *flies* is temporarily more probable than the eventual winner, *slide*. Again, this flip-flopping between roughly equiprobable hypotheses means that a decoder with a not too cautious choice of ε would be in a high danger of committing an error of type II.

This oscillation between competing hypotheses also means that the online decoding comes with its own version of the "garden path" phenomenon: Hypotheses with high prior probability attract early attention which in retrospect may turn out to be unwarranted. This is, of course, in particular the case if the

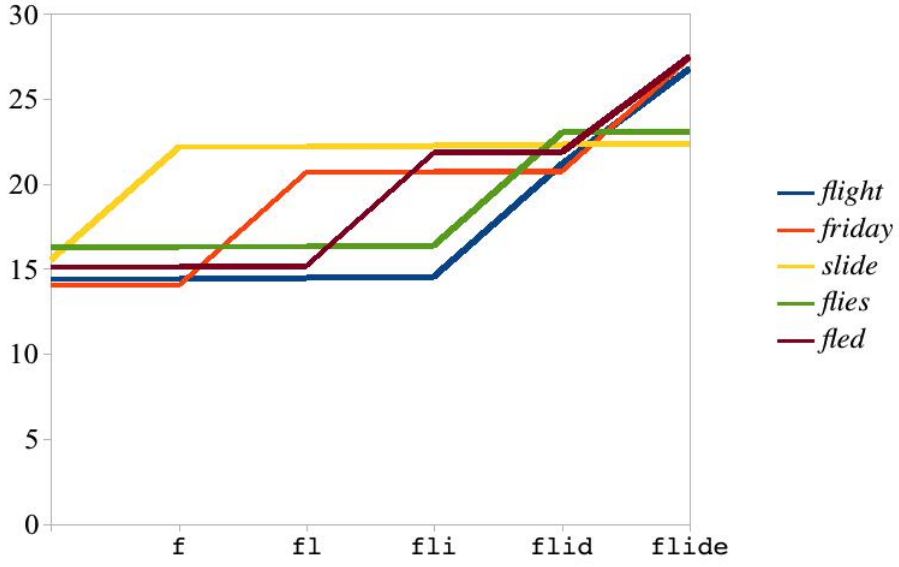


Fig. 11. Letter-for-letter surprisal values for various decoding hypotheses about the string *flide*.

output word contains errors or is otherwise abnormal so that ordinary expectations do not apply. Regardless of the cause, however, it should be emphasized that this phenomenon occurs here in a situation that does not involve any kind of "syntax" except the possibility of swapping two neighboring letters.

We have thus seen two kinds of statistical error related to two kinds of manipulated symbols string. Figuratively speaking, one abruptly pushes the receiver away, while the other keeps pushing the receiver back and forth between a portfolio of hypotheses that all seem to fit somewhat, but not quite. The first is associated with errors of type I, and the other with errors of type II.

5 Conclusion

In section 4 of this paper, I presented a simple decoding model that could distinguish between "nonsensical" and "broken" spelling — between words like *srxkz* and words like *flide*. In spite of its almost anti-grammatical nature, this model turned out to contain quite remarkable equivalents of psycholinguistic phenomena like garden-pathing, which is usually associated with the processing of context-free languages.

Because the model is probabilistic, it also contains information-theoretical tools for precisely defining what it means for an observed string to be "surprising"

and "meaningless." In section 3, I articulated those concepts in terms of a notion of statistical error related but not completely identical to the classical notion of type I and type II errors.

It is my claim that these two concepts map onto the N400 and the P600, respectively. The utter surprise associated with the N400 should be understood as a reaction to signals that have no plausible interpretations. The P600 should be understood as a correlate of the cognitive effort involved in understanding something that could be interpreted in several conflicting ways. This effect is presumably particularly pronounced when the various hypotheses swap places in the probability ranking when more data comes in. I have tried to argue for the plausibility of this interpretation by working through a number of specific examples and show what the mathematical predictions were.

The way I presented the decoding model used in this paper, its main purpose was detecting and correcting misspellings. From the perspective of the model itself, however, any other random process unfolding linearly in time would do just as fine. The individual tokens fed through the communication channel could equally well be words, coin flips, horse kicks, and letters.

Interestingly, however, there are empirical indications that many of the phenomena discussed in section 2 also work at the level of actual spelling. A study in Dutch by Vissers et al. [2006] found differences between the brain responses to misspellings depending on context:

- In die bibliotheek lenen scholieren boekun om ...
  ("In the library, the pupils borrow bouks to ... " — P600)
- De kussens zijn opgevuld met boekun waardoor ...
  ("The pillows are stuffed with bouks so that ... " — N400)

This effect seems like a good candidate for an explanation in terms of decoding. More interesting, however, it also suggests that the P600 and the N400 may not be characteristic of any single linguistic phenomenon, but rather of something having to do with the way people make sense of their experience as such. This would be consistent with the claim that the two types of brain response are connected to general processes of statistical inference rather than with some highly purpose-specific syntactic parser.

If this is true, it could have some dramatic consequences for linguistic theory. We might have to reopen the discussion that Chomsky essentially closed down half a century ago when he convinced everybody that there existed a meaningful notion of a "grammatical sentence" which has nothing to do at all with making sense. But perhaps "making sense" was exactly the keyword that we should have been paying attention to; and perhaps we already have all the mathematical tools necessary for writing a good theory of what it means to make sense or fail to make sense of the world around us.

It could also be that I am just wrong. There is always a serious danger that we come to see the whole world as nails when we have a hammer, and I just happen to have information theory in my hand instead of Chomskyan linguistics. And there is certainly something about mapping two brain phenomena onto a couple of theoretical concepts from statistics which seems just as suspiciously

neat and easy as mapping them onto two concepts from linguistics. I recognize this suspicion as a legitimate concern, and I have nothing to say against it in its general form.

What I can say, however, is that two dogmas are better than one. Linguistics has a tendency to internalize its own concepts to such a degree that they almost seem to dissolve into the phenomena, and I think the time is ripe for a fresh discussion of some of them. Going back to the very beginning seems like a good way to start.

Bibliography

Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

Noam Chomsky. *Syntactic Structures*. Walter de Gruyter, The Hague, 1957.

T Cover. An achievable rate region for the broadcast channel. *IEEE Transactions on Information Theory*, 21(4):399–404, 1975.

T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, 1991.

Fred J Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

W. Nelson Francis and Henry Kucera. *Computational analysis of present-day American English*. Brown University Press, Providence, RI, 1967.

John C. J. Hoeks, Laurie A. Stowe, and Gina Doedens. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19(1):59–73, 2004.

Albert Kim and Lee Osterhout. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2):205–225, 2005.

Herman H. J. Kolk, Dorothee J. Chwilla, Marieke van Herten, and Patrick J. W. Oor. Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, 85(1):1–36, 2003.

Gina R Kuperberg. Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49, 2007.

Gina R. Kuperberg, Tatiana Sitnikova, David Caplan, and Phillip J. Holcomb. Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1):117–129, 2003.

Marta Kutas and Steven A Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

Jerzy Neyman and Egon S Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20A(1/2): 175–240, 1928.

Mante S. Nieuwland and Jos J. A. Van Berkum. Testing the limits of the semantic illusion phenomenon: Erps reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3):691–701, 2005.

Lee Osterhout and Phillip J Holcomb. Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. *Language and Cognitive Processes*, 8(4):413–437, 1993.

Frank Rösler, Peter Pütz, Angela Friederici, and Anja Hahne. Event-Related Brain Potentials While Encountering Semantic and Syntactic Constraint Violations. *Journal of Cognitive Neuroscience*, 5(3):345–362, 1993.

Edward Walter Samson. Fundamental natural concepts of information theory. Technical Report E 5079, Air Force Cambridge Research Center, 1951.

Claude E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27:379–423, 1948.

Marieke van Herten, Herman H. J. Kolk, and Dorothee J. Chwilla. An erp study of p600 effects elicited by semantic anomalies. *Cognitive Brain Research*, 22 (2):241–255, 2005.

Marieke van Herten, Dorothee J. Chwilla, and Herman H. J. Kolk. When heuristics clash with parsing routines: Erp evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, 18(7):1181–1197, 2006.

Constance Th WM Vissers, Dorothee J Chwilla, and Herman HJ Kolk. Monitoring in language perception: the effect of misspellings of words in highly constrained sentences. *Brain Research*, 1106(1):150–163, 2006.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13 (2):260–269, 1967.

Jill Weckerly and Marta Kutas. An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology*, 36(5): 559–570, 1999.