NATURALNESS IN LANGUAGE**

JOHN SINCLAIR
Department of Language and Literature - University of Birmingham

The argument in this paper is that there exist a very large number of well-formed sentences which do not seem natural to a sensitive native speaker; therefore these sentences must violate some restrictions which are not among the criteria for well-formedness. It is important to examine what the further restrictions might be, for at least three reasons:

a. there is no reason to believe that the restrictions are any less central in language structure than those for well-formedness;

b. decisions about well-formedness are normally made on sentences in isolation, by people whose intuitions are shaped by experience of continuous text. As text study grows in importance, the concept of well-formedness of sentences in text may prove to be of value. This concept I want to label naturalness, for the time being;

c. the concept of naturalness may be particularly useful to the learner of a language.

Sentences which are not natural do occur. Typically they occur in text composed for purposes other than direct communication, and in the speech and writing of non-native speakers. (I reserve for a future occasion the discussion of whether or not native speakers can produce non-natural utterances in normal situations.)

(1) Prince Charles is now a husband
(2) I'm trying to rack my brains.

Ex. 1 is quoted from a language-teaching coursebook; ex.

This article was first published in **Corpus Linguistiis Aarts J. and Meys W. (eds.) 1983, p.203-10.

2 from the speech of a fluent speaker of English as a second language. My claim is that they are not natural, though they appear to be well-formed.

The professional linguist has to make some adjustments in order to appreciate the distinction between well-formedness and naturalness, because it is a well-established convention that sentences like ex. 1 are composed for purposes of exemplification; the linguist's tolerance of non-naturalness is abnormally high. There is even a point of view that assumes that a context must exist for every well-formed sentence and much ingenuity goes into attempts to ignore the criterion of naturalness.

It might be supposed that naturalness will always be probabilistic, and therefore distinct from well-formedness, which is absolute. Certainly the textual evidence for naturalness is probabilistic to begin with, but when sentences are described in their textual environment (or co-text) there may well be absolute or nearly absolute statements to be made about their priviliges of occurrence.

The point may be illustrated by the example of the word matter, used as a verb. Nearly all its occurrences in text are in the simple negative form — it doesn't matter, etc. A tiny number occur in the positive and/or continuous form; it should matter, his decisions have a habit of mattering. From this it would be dangerous to conclude that the verb can be used without restriction, because the nontypical forms are clearly contrastive with an occurrence of a regular form in the preceding discourse. Just as a nontypical pronunciation can occur in a contrastive context (no it was his grandfather who dies), non-natural structures can occur in text without attracting licence to appear without restraint.

This is a classical linguistic argument, transposed into the consideration of text structure. Most attempts to study text have quickly abandoned the normal rigour of sentence grammar because of the apparent freedom of choice in text;the

supposition in this paper is that text is much more determined than is normally supposed. But we are still woefully short of descriptions that marshall the available evidence, and have to be content with interim statistical statements prior to proper analysis.

When a sentence is examined in isolation, the judgement about naturalness is sensitive to some matters, and not to others. Taking the latter first, there are some requirements of text which do not need to be realised in any particular sentence, and to not affect well-formedness. For example, a third person pronoun requires a referent — from the surrounding text, the immediate situation or the shared knowledge of participants.

(3) I wouldn't have bought it if he hadn't been there

The absence of referents to it and he do not disturb the impression of naturalness; similarly there.

(4) I wouldn't have bought it if I had known

The verb know normally requires an element to follow it, either a report (... that that exchange rate is poor) or a noun phrase (the exchange rate). Example 4 is quite natural but the missing expression is clearly a report.

Another example of what I call allowables is the restriction on the occurrence of past tense verbs. A simple past tense in English must occur in the environment of a point-time referent, if it is to share the meaning of an event in the past. The time referent does not need to be in the same sentence, because such items are "written down" from sentence to sentence until superseded. But unsupported past tenses are commonly quoted as quite natural:

(5) The cat sat on the mat

There is another set of choices which is not so readily accepted. The sentences which contain them are heavily text-dependent because in isolation they might not even be considered well-formed. But they occur in profusion.

To illustrate this set, I would like to return to the contrast between well-formedness and naturalness and suggest that they are independent variables; thus

	well-formed	ill-formed
natural	ex. 6	ex. 7
non-natural	ex. 8	ex. 9

Figure 1

(6) I am writing this paper for you to read.
(7) If you like.
(8) Look forward to clapping eyes on you.
(9) Book the.

Example 7 is a natural response to a particular kind of request, but it is not well-formed by regular rules of English sentence structure. One might argue that it is an idiomatic expression which is largely fixed in form, but many such idioms are well-formed. Also, there are many examples of my second type which are not particularly idiomatic, eg

(10) The goalkeeper kicked.

Wherever a normally transitive verb lacks an object the sentence will not be considered well-formed.

This second type of sentence includes choices which indicate text dependency of various kinds, and these I shall call rangefinders. The reason for the name is that somewhere in the cotext or the context will be found the item signalled by the dependency choice — or the text is problematic. It should be possible to discover the range of text required in order to cover the dependencies that are signalled by a sentence.

Some of these choices show gross dependency, such as ex. 10, and are considered ill-formed as sentences; others are less obvious and form a continuum with allowables at the far end.

(11) We searched

(12) We searched all night

(13) We searched all night for the missing climbers.

In my view, ex. 12 is the most natural of the three. Ex. 11 is heavily text-dependent, and ex.13 is, so to speak, not dependent enough, and is typical of a sentence made up as an example.

There is a third set of choices which are relevant to naturalness, and these I shall call supporters. Many linguistic choices have a tendency to co-occur with each other, and so the presence of one is valuable evidence for the existence of another.

Lexical studies have long recognised the feature of collocation, though co-occurrence of syntactic choices is not so well documented. Firth proposed colligation for what appears to be lexicosyntactic co-occurrence and some modern uses of collocation are lexicosyntactic rather than purely lexical.

Supporters may be lexical,syntactic or a combination of the two. Where each word can have several mennings, ambiguity is inherent in text, and the tendency of features to co-occur is a valuable aid to interpretation.

Consider the sentence:

(14) couldn't get through all night.

To most competent users of English, this sentences immediately suggests a failed attempt at telephone contact. Yet no word in it is of strong lexical significance; in fact only night would normally be regarded as a vocabulary word. The phrasal verb get through has nany meanings, of which "make telephone contact" is not likely to be a very frequent one. What, then, are the clues that suggest the immediate interpretation?

One important clue is that the clause is intransitive. Another is the selection of all night, which would not go

comfortable it get through referred to physical progress. We must note

a. that a time adjunct has been selected
b. that it refers to a period of time, and
c. that the particular item all night occurs.

These features, then, if they were demonstrated to contribute substantially to the interpretation, would be supporters. But on this occasion, possibly, features such as negative modal could, pronoun I do not seem to be germane. Statements about which features are supporters are deliberately cautious in the present state of our knowledge.

The analysis of a sentence will be in terms of **allowables, rangefinders** and **supporters** in the first instance. Allowances will be made for features of register and other types of systematic variety and for shared knowledge and experience. Then some observations will be made about the naturalness of a sentence.

The form of naturalness statements is currently in terms of three parameters:

- neutrality
- isolation
- idiomaticity

A very neutral sentence is one which has few if any mutual support choices, and can be cryptic, vague, or trivially ambiguous. Naturalness occupies a middle band, and at the opposite end from neutral is a sentence which is too intricately constructed to sound natural, as one may find in a mannered literary style.

A fully isolated sentence is one which contains no rangefinders or allowables. At the other end of the scale is a very heavily text-dependent sentence, which in isolation would be judged ill-formed.

An idiomatic sentence is one in which the evidence of allowables, supporters and rangefinders does not conflict, and

where there is sufficient evidence to sustain variation. Conversely, an unidiomatic sentence will either contain conflicting evidence or apparently unmotivated variation.

We can return to earlier examples to see this classification in action.

1. **Prince Charles is now a husband**

   neutrality: low; note that a supporter such as good modifying husband would immediately improve the naturalness of the sentence.
   isolation: extremely high
   idiomaticity: low; there is a conflict between the mutual expectations of the equative structure, the indefinite article, and the word husband. Words denoting occupations (eg sailor) would not cause this conflict.

2. **I'm trying to rack my brains**

   neutrality: medium
   isolation: rather low; rangefinders concern what purpose to be served by the mental activity.
   idiomaticity: very low; conflict between expectations of I'm trying and rack my brains.

3. **The cat sat on the mat**

   neutrality: medium
   isolation: fairly high. There is the Simple Past tense requiring a point-time adverbial and two instances of the; all allowables.
   idiomaticity: medium, but difficult to assess because of the common use of this sentence as an example in language work.

4. **I am writing this paper for you to read**

   neutrality: medium; strong collocations write, paper, read

isolation: very isolated. All three nominals are self-identifying, and the only oddity is the paradox induced by giving the sentence an example number

idiomaticity: medium

5. It should matter

neutrality: the missing negative would be a strong support; the particular sense of it is supported by the word-class of matter

isolation: heavily text-dependent, because of the strongly expected supporting negative, probably in the previous sentence

idiomaticity: very low because of no textual evidence to sustain gross variation

I shall try to illustrate these three parameters further by making changes to a sample sentence. Ex. 16 is verbatim from text.

(16) Each had accumulated large reserve stocks before the war.

To make this sentence more neutral, we could replace accumulated by built up, or even delete it altogether. Large could become big, reserve could vanish, and stocks could become quantities. Before the war could be replaced by a vaguer time reference like earlier. Figure 2 illustrates the stages.

NEUTRALITY

Each had	accumulated	large	reserve stocks	before the war
Each had	built up	large	reserve stocks	before the war
Each had	built up	big	reserve stocks	before the war
Each had	built up	big	stocks	before the war
Each had	built up	big	quantities	before the war
Each had	built up	big	quantities	earlier
Each	built up	big	quantities	earlier

ISOLATION

Each had accumulated large reserve stocks before the war
Each had accumulated large reserve stocks of paper before the war
Each accumulated large reserve stocks of paper before the war
Each accumulated large reserve stocks of paper before 1939
Printers accumulated large reserve stocks of paper before 1939

IDIOMATICITY

Each had accumulated large reserve stocks before the war
Each before the war had accumulated large reserve stocks
Each before the war had large accumulated reserve stocks
Each before the war had large accumulated reserved stocks
Each before the war had mighty accumulated reserved stocks
Each before the war had mighty reserved accumulated stocks
Each before the war held mighty reserved accumulated stocks
Each before the war held mighty reserved accumulated stores
Each before war held mighty reserved accumulated stores

Figure 2

To make sentence 16 more isolated, we could add in what type of stocks were accumulated, eg of paper; remove had, and give a date to the war, eg 1939. Each would have to go, replaced by, perhaps, Printers.

To make sentence 16 less idiomatic, we could reposition before the war to a less usual place, and turn accumulated into a modifying participle. Changing reserve to reserved clogs it up still more, and a change from large to mighty destroys collocations. The is deletable but clumsy, had could be replaced by held, and stores could replace stocks. Fig. 2 shows what an unidiomatic version of the sentence would read like.

Not all of these changes will be agreed by everyone as effecting a move on the relevant parameter, or in the direction intended, or maintaining a perception of well-formedness in structure. I have not attempted to illustrate moves in the

other directions on the three parameters but it is clear that deletion of the time adjunct will decrease isolation, and a more technical phraseology might decrease neutrality further. I would not attempt to make the sentences more idiomatic.

To proceed further with text structure at sentence level, we need to make certain assumptions that can be checked against a large body of data. From the study of structure below the sentence we take over three assumptions.

a. Structure is realised by recognisable signals in text. Not every meaningful choice is unambiguously realised, but the statement would be generally agreed.

b. There are a number of non-trivial structural statements which entail a high level of generality. Although some structural statements can be made about small details, there is a high value placed on the generalities — they are called "powerful" etc.

c. Structure is predictive in that not all possible combinations of elements can occur; therefore each selection of an element reduces the number of options that remain.

To these we must add:

d. A stronger version of (a) — each distinct meaning in a structure is associated with a distinct pattern of choice involving more than one item. "Associated with" can be understood probabilistically at this stage.

e. The same repertoire of signals is used many times, so that general statements can be made without loss of contact with the data.

These assumptions must now be tested through an extended study of texts, which will establish the precise conditions for naturalness. The study of allowables will lead to the specification of an abstract text framework for any sentence. The study of rangefinders will

a. show how each sentence is integrated into its text
b. establish the range of individual features.

The study of supporters will tell us a lot about the resolution of textual ambiguity, and will lead to a precise specification of

a. complex items, eg phrases

b. permitted range of variation.

The three scales of neutrality, isolation and idiomaticity will allow sentences to be compared with each other and might lead to a modern rhetoric at the rank of sentence.

The general concept of the well-formedness of text is arousing interest at present, and naturalness is offered as a useful category to describe textual well-formedness among sentences.