THEORETICAL SYNTAX MEETS THE WORD-COUNTERS

Professor Richard Hudson, FBA
Department of Phonetics and Linguistics
University College London

Grammatical research has tended for a long time to have two very different focuses: the structure of the grammatical system itself, and the grammatical structure of texts. Those interested in the system have seen little point in studying texts, as their main source of data has been native-speaker intuitions; and conversely, new ideas about how grammatical systems work have had little interest for text-students. This paper, by a theoretical syntactician, will review some recent work on texts which seem to raise interesting theoretical questions.

1. Word-class frequencies. The proportion of words that belong to any given word-class turns out to be much more constant across texts than one might expect. The theoretical question is ‘Why?’, but no answer is even on the horizon. Moreover, if we compare genres we find very small but constant variations in these proportions; and likewise if we compare languages. By a fortunate chance, one of the sources of data is the Greek New Testament, in which the word-class frequencies are astonishingly similar in some respects to those of modern English, though interestingly different in other respects. These data (and others) suggest a new approach to language typology, as well as to genre comparisons and even to sociolinguistics.

2. Word-order and dependencies. Suppose we make the controversial theoretical assumption that the foundation for sentence-structure is a surface dependency analysis (linking words in pairs, without intervening phrases). One practical consequence is that it becomes easy to analyse large texts syntactically (and to teach this skill to students), which in turn means that it becomes possible to ask some simple questions about the relative frequencies of various patterns in the dependency structures. How many words are more than 1 (or 2 or ...) words from the word on which they depend (their head)? How many words precede (or follow) their head? How many adjectives depend on a following noun? The answers to these questions may be relevant to theoretical questions ranging from questions about how children learn language to questions about relative markedness of constructions.

3. Syntactic complexity. One way to measure syntactic complexity is to count the number of ‘incomplete’ dependencies after each word, as each such dependency places an extra burden on short-term memory. It is easy to measure the syntactic complexity of a whole text in this way, and to compare different texts (even across languages). This measure may also explain why it is so hard to understand self-embedded sentences like The dog the cat the farmer bought chased barked.
Theoretical syntax meets the word-counters in Greek and English

Richard Hudson
University College London

1. Introduction

This paper is an attempt to show that word-counting can provide important evidence for descriptive and theoretical linguists. This is a fairly unusual claim, but not a controversial one; only a minority of linguists count words, but so far as I know no-one has ever gone so far as to say that word-numbers are always illegitimate as evidence.

The most acceptable role for quantitative evidence is in distinguishing clearly defined hypotheses (Smith 1989:184). It is not hard to think of linguistic analyses that are clear enough to generate numerical predictions. For example, suppose three linguists are working on a language like Greek, in which the order of clause elements is free; in particular, a verb and its subject and object can occur in any of the logically possible orders VSO, VSO, VOS, SVO, OVS, SOV, OSV.

Linguist A claims that order really is free, i.e. it is not controlled by anything at all, whether syntactic rules based on grammatical relations, or semantic rules based on semantic roles, or even pragmatic matters such as topic, weight, definiteness and so on. According to this theory the choice of order is completely random, just like the result of shaking a dice. Just like each of the six numbers on a dice, each has an equal chance of being used on any given occasion. I can't imagine any linguist actually espousing such a theory, but at least it is clear and easily testable, because it predicts that in any text the six orders should all be represented equally (subject to the small amount of variation which is to be expected by chance).

Linguist B is convinced that the order of elements is completely free as far as syntax is concerned but is heavily influenced by pragmatic factors. There are a host of different pragmatic factors that could be relevant, but let's suppose that for this particular linguist, the most important question is the idea of 'information flow' (Chafe 1992). In this theory, what counts is the relative 'newness' of the elements, or rather of their referents, and elements with referents that are new, i.e. not already known to the addressee, follow those with 'old' referents. It doesn't matter if the notion of newness is rather vague; if there is any truth at all in it, a statistical analysis based on the crudest of distinctions should reveal some kind of effect. For instance, if one of the factors contributing to newness is
definiteness and noun phrases can easily be classified for definiteness, then we can measure definiteness, as an imperfect approximation to the vaguer notion 'newness'. If newness really is an important determinant of order, the percentage of noun phrases which are definite should decrease as we move through the clause. Each of the six possible orders including S and O should show more definiteness in the first of these noun phrases than in the second (including the orders VSO and VOS).

Linguist C agrees with B that newness has something to do with the order of elements, but thinks the syntax also plays an important part. The part played by syntax is actually rather surprising, because this linguist thinks that the syntax makes the position of S depend on that of O and the position of O depends on whether or not it is 'focussed' (a notion which can fortunately be left more or less undefined for present purposes). Specifically, if O is focussed, it must precede V, and if O precedes V, S must follow V - there is a strong preference for S and O to stand on opposite sides of V. Otherwise the position of S, before or after V, is more or less random. This theory, like the others, allows various statistical predictions as we shall see later.

Once we know how frequent the various orders are in freely occurring texts, it should be possible to test these theories and eliminate some of them. Fortunately some very useful statistics on Greek are in fact available, and we shall return to them later.

There is no doubt that statistics of language use are good for testing fully-fledged theories; but this is not their only use. The figures themselves can reveal completely unexpected patterns which call for a theoretical explanation. Just as the movements of the observable stars allow astronomers to predict the existence of unseen heavenly bodies (or even of invisible matter), so the statistical patterns that we can discover in texts allow us to predict the existence of influences on the choice and use of words even when we have no clear idea what these influences are. In some cases, we may be lucky: the predicted influence may be one we already know about. In such cases we can celebrate the fact that a theory based on one kind of data turns out to explain a completely different kind of data. Deduction has met up with induction, and the two have reinforced each other. But in at least some cases no theoretical explanation is available. The statistical pattern is no less real for being unexplained, nor for being statistical. The next section describes an example of this kind, which again is relevant to Greek.
2. Word-class frequencies

What follows is an example of pure inductive research (Hudson 1993, 1994). I was curious to know whether there was any regularity at all in word-class frequencies, i.e., in the proportion of words in a text that belonged to each of the major word-classes. I started without any theoretical expectations at all, and was certainly not ‘testing’ any theory. I started with some very short texts (of a mere 100 or so words), and was surprised to find that the percentages were rather constant - verbs clustering round 18%, adjectives round 9%, and nouns round 37%. The range of variation was quite small, but rather surprisingly it was less for nouns than for the smaller classes. Why should there be any regularity at all in the percentage of word-classes, and why these particular figures?

Having started on a hunt like this it is very hard to stop. How general are these patterns? What about other kinds of English? What about other languages? What about children’s language? And so on and on. The questions ask themselves, and all the researcher has to do is to try to answer them. The researcher may have a more theoretical agenda as well; I, for one, would dearly like to know why the patterns occur, but I’m certainly not going to delay the fun of finding new patterns until I have some testable theories that might explain them. Meanwhile we can already eliminate some theories from our enquiry, as I shall explain below.

To cut a long story short, I have found massive support for these first tentative conclusions. The proportion of words contributed to a text by the major word-classes does indeed tend to be remarkably constant, and the more precisely we define the kind of text, the more constant the word-class frequencies are. Two very large corpora of written English have been analysed in terms of grammatical word-classes, and the results have been published (Francis and Kuera 1982, Johansson and Hofland 1989). The first is the million-word Brown corpus, and the second, also a million words long, is the Lancaster-Oslo-Bergen (LOB) corpus. Table 1 shows the percentage of words that belonged to various word-classes, first for the two entire corpora, and then for their two major subdivisions, called ‘informational’ and ‘imaginative’ (i.e. roughly fact and fiction).
<table>
<thead>
<tr>
<th>word-class</th>
<th>whole corpus</th>
<th>informational</th>
<th>imaginative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOB</td>
<td>Brown</td>
<td>LOB</td>
</tr>
<tr>
<td>NOUN</td>
<td>35.7%</td>
<td>37.3</td>
<td>35.9</td>
</tr>
<tr>
<td>common/proper</td>
<td>26.9</td>
<td>28.8</td>
<td>29.1</td>
</tr>
<tr>
<td>pronoun</td>
<td>8.8</td>
<td>8.5</td>
<td>6.8</td>
</tr>
<tr>
<td>VERB</td>
<td>17.7</td>
<td>18.2</td>
<td>16.4</td>
</tr>
<tr>
<td>PREPOSITION</td>
<td>12.1</td>
<td>12.0</td>
<td>13.1</td>
</tr>
<tr>
<td>AD-WORD</td>
<td>12.7</td>
<td>12.3</td>
<td>12.8</td>
</tr>
<tr>
<td>adjective</td>
<td>7.2</td>
<td>7.1</td>
<td>7.8</td>
</tr>
<tr>
<td>adverb</td>
<td>5.5</td>
<td>5.2</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Table 1: Major word-class frequencies in the LOB and Brown corpora of written English and their subdivisions.

Perhaps the most striking thing about the figures in this table is the constancy in the figure for all nouns - i.e. for common or proper nouns combined with pronouns. For each corpus, the contribution of nouns is almost exactly the same in both the sub-corpora, with only 0.2% or 0.3% variation - i.e. a mere 2 or 3 parts in a thousand. Admittedly the two major corpora are rather different, with Brown consistently about 1.5% higher than LOB, but this is presumably due to some small difference in the way words were classified. Why should the representation of nouns be so constant? And why these figures (36 or 37%) rather than some quite different figures?

What makes the constant figures for all nouns even more interesting is that it masks a considerable variation between sub-genres in the balance between pronouns and other nouns. Informational texts contain far more common or proper nouns than imaginative texts do, but far fewer pronouns, with differences of nearly 10% in some cases. Why should informational and imaginative texts differ in this way?
New Testament are divided into letters and narratives. Computer analysis of text does figures are shown in Table 2, where the books of the New Testament from the University of Pennsylvania, Center for the Australian Understanding, James Tarter, who had obtained a grammatically 'fitted' copy of the New Testament, so may not apply to modern Greek. They were supplied to me by an

figure on Cretan, but unfortunately they are based on the Greek text of the New Testament. They may not apply to modern Greek. They were supplied to me by another kind of theoretical question is in which context these figures are due to the peculiarities of English grammar, as opposed to general, perhaps universal, facts about language use. So what about other languages? By a rather happy coincidence, I have some

Also, by definition, a language is a set of symbols that can be stored, and recovered.

Some common nouns are given in Hudson (1993). This, then, is an example of

someone and nothing are more common in imaginative writing, and correlate negatively to

tendency affects all pronouns (except who/whom). Even compound pronouns like

pronoun. Interestingly, it turns out that this is in fact a blind alley, because the same

pronoun which can easily be answered if we have data on the distribution of other kinds of

differences in the use of personal pronouns. This is the interesting, statistically

explanation further we have to explain why the statistical differences are found in

more common in fiction than in factual writing. However, before we pursue this kind

of noun?. If so, the question changes, and we have to explain why personal pronouns a

and so on, but of course a pronoun is an alternative to a description in terms of com-

novels use words like I and she more often than newspaper editorials, scientific articles

empirical to explain this in terms of the use of anaphoric and personal pronouns. May

conclusions. At least we may be in a position to explain some of the differences. It is

This is a theoretical question, and even if we can't explain the simultaneity and
<table>
<thead>
<tr>
<th></th>
<th>whole New Testament</th>
<th></th>
<th>narrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>32%</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>common/proper</td>
<td>21</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>pronoun</td>
<td>12</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>VERB</td>
<td>20</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>PREPOSITION</td>
<td>8</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>AD-WORD</td>
<td>12</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>adjective</td>
<td>7</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>adverb</td>
<td>5</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2: Major word-class frequencies in the Greek New Testament and in its two divisions (letters and narrative).

Let's consider first the pattern of nouns that we found in English. Once again the overall percentage of nouns is virtually the same in the two subdivisions, and once again this conceals a major difference in the balance of common or proper nouns to pronouns, with far more common or proper in the letters, but fewer pronouns. This repeats the pattern of English, provided we can identify 'letters' with the 'informative' category and 'narrative' with the 'imaginative' one (a move which I, for one, find reasonable). It's true that the actual figures for Greek are somewhat lower than the corresponding English figures, and that this calls for an explanation; but the trend is clearly the same in both languages.

Another trend in both languages involves verbs and adjectives. In this case, even the figures are the same (if we round up or down to the nearest percentage point). In both languages, 17% of words are verbs in informational/letter texts, but 22% in imaginative/narrative texts; and 8% are adjectives in the first kind of text compared with 6% in the second kind. Even prepositions show a similar trend, with both languages showing more prepositions in the first kind of text than in the second kind.

These data point to a general explanation of the genre differences which cuts across differences between languages, an explanation in terms of the different communicative functions that members of different word-classes perform. Such an explanation must be linguistically sophisticated, and in particular it must not fall unthinkingly into the trap of
assuming a simple relationship between word-classes and semantic function. The same
kind of thing can often be referred to equally well either by a noun or by a verb (e.g.
explosion or explode), so it’s useless to look for an explanation in terms of how often we
need to refer to events, people, concrete objects, and so on. Another fact about the
numbers for Greek and for English is that prepositions are more common in English texts
than in Greek ones. This is easy to explain functionally, given that Greek uses inflectional
cases to express a lot of information which requires a preposition in English. However it
would be comforting if we had an analysis of each language which would allow us to
predict the precise numbers found, and the difference between them, but this we don’t
have at present.

This discussion certainly raises far more questions than it answers. What about
Modern Greek? What about spoken Greek, and spoken English? What about other
languages? What about children’s language - is there a developmental scale in the use of
word-classes? How regular and robust are these figures? All these questions are simply
matters of observable fact, which we already know how to answer once we have access to
the right kind of material; and indeed tentative answers are already available for some of
them (Hudson 1994). These questions are just the foothills, of course, and the really hard
part of the climb only starts when we look for explanations. Why are the constant figures
constant, and why do some other figures vary? And perhaps most challenging of all, why
these numbers, rather than higher or lower ones? When we can answer that question we
really will have something to be proud of.

3. Word-order in Greek

I want to illustrate the benefits of a statistical approach in the other area of syntax that I
have already discussed, word order; and specifically, I want to discuss the order of the
clause elements Subject, Verb and Object in Modern Greek. My data comes from two
sources: Dimitra Tzanidaki, a research student in our department who is working on this
problem from a more theoretical point of view, and Dr Chryssoula Lascaratou, who
analysed a large amount of written material in Modern Greek and published her figures
together with a very interesting theoretical discussion (Lascaratou 1989). Ms. Tzanidaki
and I have reworked some of Dr Lascaratou’s data, so what I am reporting here is part of
an ongoing joint project.

The starting point for the discussion is the fact (noted above) that all of the six
logically possible orders of S, V and O are possible. For example, all the following sentences mean 'Peter met Eleni', with sinantise meaning 'met':

(1) sinantise o petros tin eleni.  VSO
(2) sinantise tin eleni o petros.  VOS
(3) o petros sinantise tin eleni.  SVO
(4) o petros tin eleni sinantise.  SOV
(5) tin eleni o petros sinantise.  OSV
(6) tin eleni sinantise o petros.  OVS

What is the theoretical status of these alternative orders? In particular, how do speakers choose between them? Let's start by eliminating the extreme answer of linguist A which I mentioned earlier, namely that the choice is simply random, like tossing a dice. As I suggested in the earlier discussion, this is easy to test statistically because it predicts that each order should have the same chance of occurring. The facts certainly don't support this theory, as can be seen from Table 3, based on the figures given in Lascaratou (1989: 42).

<table>
<thead>
<tr>
<th>Order</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVO</td>
<td>1246</td>
<td>80.9%</td>
</tr>
<tr>
<td>OVS</td>
<td>219</td>
<td>14.2%</td>
</tr>
<tr>
<td>VSO</td>
<td>27</td>
<td>1.7%</td>
</tr>
<tr>
<td>SOV</td>
<td>19</td>
<td>1.2%</td>
</tr>
<tr>
<td>VOS</td>
<td>19</td>
<td>1.2%</td>
</tr>
<tr>
<td>OSV</td>
<td>10</td>
<td>0.6%</td>
</tr>
<tr>
<td>Total</td>
<td>1540</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 3: Frequencies of six orders of S, V and O in Modern Greek written texts.

The second theory that I mentioned was that of linguist B, who believed that order was influenced by the tendency for 'given' information to precede 'new'. Assuming that definiteness is closely related to givenness, this predicts that where a clause contains two NP's, the first is more likely to be definite than the second, a prediction which should be very easy to test. Unfortunately the available data don't address this particular question, but they do allow us to approach the more general question indirectly by exploring the
relative definiteness of subjects and objects in different positions (Lascaratou 1989:55-62). The figures give only partial confirmation of the prediction. As far as objects are concerned, the trend is very clear. If we start with the simplest cases, where there is an object but no subject at all, only two orders are possible: V O and O V. The figures are shown in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>definite O</th>
<th></th>
<th>indefinite O</th>
<th></th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>VO</td>
<td>481 (63%)</td>
<td></td>
<td>278 (37%)</td>
<td></td>
<td>759 (100)</td>
</tr>
<tr>
<td></td>
<td>(83%)</td>
<td></td>
<td>(98%)</td>
<td></td>
<td>(88%)</td>
</tr>
<tr>
<td>OV</td>
<td>102 (94%)</td>
<td></td>
<td>6 (6%)</td>
<td></td>
<td>108 (100)</td>
</tr>
<tr>
<td></td>
<td>(17%)</td>
<td></td>
<td>(2%)</td>
<td></td>
<td>(12%)</td>
</tr>
<tr>
<td>total</td>
<td>583 (67%)</td>
<td></td>
<td>284</td>
<td></td>
<td>867 (100)</td>
</tr>
<tr>
<td></td>
<td>(100)</td>
<td></td>
<td>(100)</td>
<td></td>
<td>(100)</td>
</tr>
</tbody>
</table>

Table 4: Definiteness of O by position of O in VO and OV clauses.

However you look at Table 4, it confirms the prediction that definiteness favours an early position in the clause - 94% of preposed objects are definite, compared with the average of 67%, and 17% of definite objects are preposed, compared with the average of 12%. The trouble with this finding is that its link to the original claim is rather tenuous, because we don't know anything about the newness of the verb. The information in a verb may presumably be more or less new, just like that in a noun, but we don't know whether the object in an OV clause is more or less new than the verb. Even if relative givenness predisposes an object to precede the verb, we don't know whether this is because it is more 'given' than the verb. Without an independent way of measuring the givenness of the verb, we can't test the hypothesis.

The picture is even less clear when we consider clauses that contain a subject as well as an object. (For the sake of simplicity I am deliberately restricting the discussion here to active clauses, although Lascaratou gives a great deal of useful information about passives.) Lascaratou's figures are shown in Table 5; unfortunately she doesn't give figures for SOV and OSV, and some cases seem to be missing from this count. (I suspect the figures in this table exclude the figures for clausal objects.)
Table 5: Definiteness of O by position of O in clauses containing S.

Let's start by comparing the two really large order-categories, SVO and OVS. The figures presented confirm the tendency for definite objects to be preverbal. Reading down the columns, we find that a higher proportion of definite objects than of indefinites are found in OVS (20% compared with 12%), and conversely a lower proportion are found in SVO (76% to 86%). Moreover, reading horizontally, the object is less likely to be definite in an SVO clause than in an OVS clause (63% to 76%). One of the smaller categories (VSO) confirms this trend, while the other (VOS) conflicts with it. But maybe we should ignore these figures, since the totals are very tiny.

Unfortunately, all that these figures show is what we have already found, namely that definite objects tend to be positioned before the verb. They tell us nothing about the relation between the object and the subject. What about subjects? Do they show a similar tendency to be positioned according to their definiteness? The answer, according to Lascaratou (p. 56) is no: "in the active we cannot detect any correlation between subject [+/− definiteness] and surface structure orderings." Lascaratou argues that the statistical tests may be affected by the fact that active subjects are overwhelmingly definite (91% definite, in fact), but I'm not convinced. If there had been a trend, her statistical test (the Chi-square test) would have shown it.

I therefore prefer to conclude that definiteness is irrelevant to the positioning of subjects; and if this is so, then it is also irrelevant to the positioning of subjects in relation to objects. Linguist B seems to be wrong, though the theory contains a germ of truth, namely that definiteness is relevant to ordering. Its relevance seems to be different from
the one predicted, and concerns only the relation of the object to the verb.

The theory that we have just eliminated isn’t, of course, the only possible theory that relates word order directly to pragmatic functions. Others could involve matters of focus, for example - and indeed we shall see below that focus, like definiteness, probably is relevant to Greek order. The main point that I have tried to make is a methodological one: that it is possible to test such theories statistically, and to arrive at reasonably firm conclusions.

We now turn to the third kind of theory, one which involves questions of syntactic structure as well as general pragmatic considerations such as definiteness. This kind of theory is much easier to reconcile with the figures in Table 3, which showed the relative overall frequencies of the six orders in transitive clauses. The most striking feature of this table is the enormous difference in importance between SVO and all the other orders - 81% for SVO, with the remaining 19% shared (unequally) among the other five orders. It is at least tempting to take this dominance of SVO as evidence that this is a normal order of elements required by the grammar of Greek; and in the absence of evidence to the contrary, I shall make this assumption. There are a variety of ways in which a grammar can define one order as normal, while allowing others as exceptions; for the time being let’s simply assume that some suitable method is available, and leave these more theoretical questions till later.

Of the remaining orders, by far the most common (with 14%) is the mirror-image of SVO, OVS. Why should these two orders between them account for 95% of all the transitive clauses? Why not two orders which appear much more similar, such as SVO and VSO, which differ in only one respect (the position of the verb)? In contrast, SVO and OVS seem much more different. However it all depends on how we count differences, which depends on how we define the orders. Suppose we start from the feature that they share, which is that S and O are separated by V, then they differ in only one respect, which could be described as the position of O. In SVO, O follows V, but not in OVS. The analysis of SVO and OVS now consists of two facts: in both cases S and O are separated by V, and what distinguishes them is the position of O before or after V. These two facts cannot be stated without mentioning the grammatical functions S and O and the word-class V, so they are indisputably facts of grammar, impossible to reformulate in purely pragmatic terms.

Now we can bring back the discussion of pragmatics. What decides between SVO and
OVS? We saw above that definiteness is relevant to the position of O, but not of S, so a first guess is that it all depends on the object, with the subject taking whichever place is left for it. This is not to say that every definite object stands before the verb; we have already seen that definiteness merely increases the chance of an object being pre-verbal. In clauses without S, 17% of definite objects are pre-verbal, compared with just 2% of indefinites; and in those with S, the comparable figures are 20% and 12%. In each case, some pre-verbal objects are indefinite, and most definite objects are still post-verbal, so definiteness is clearly not the whole of the story. Rather encouragingly, the percentage of subjectless clauses that are OV is about the same as that of OVS clauses (12% versus 14%), whereas the position of subjects in intransitive clauses turns out to follow a completely different pattern, with SV and VS being equally common (Lascaratou 1989:76).

If definiteness is only part of the picture, what else is relevant to the position of the object? According to Tzanidaki, an even more important pragmatic factor is focus, which gives some absolute judgements. This unfortunately means a shift from written language to spoken, as focus is identified by intonation; so the following remarks rest on the rather dangerous assumption that evidence from speech can be applied to writing. I shall refer to the element which is picked out by intonational focus as the ‘focussed’ element, on the assumption that the intonation is used to focus the speaker’s attention on this element.

Tzanidaki and the native speakers she has consulted agree that preverbal objects must be focussed (with the exception of SOV, where either S or O must be focus). Furthermore, if O is focussed, it must be pre-verbal (with the possible exception that VOS allows O-focus, which I shall ignore because the number of VOS clauses is so small).

This virtually absolute link between focus and O position is too clear to be explained by anything but a rule of grammar, and my guess (which is no more than a guess) is that definiteness is relevant to the position of O via its contribution to the probability of an O being focussed. In contrast with the three OV orders (OVS, OSV and SOV), the three in which O follows V need not have any focussed element at all, though either V or S may be focussed in non-final positions.

In short, the order of elements is influenced by two considerations, focus and the separation of S and O. As far as focussing is concerned, the options are no focus, S-focus, V-focus or O-focus; the order of V and O must be VO unless O-focus forces OV. The separation of S and O encourages S and O to be on opposite sides of V; unlike the
first, this is a strong trend, but not absolute. This relatively precise analysis can be tested against the numbers. If it is right, we should be able to give a probability to each of the alternatives, and write a formula which combines these probabilities to predict the figures that we observed. If we know the probability of a clause having no focus or focussed S, O or V, and also the probability of S and O being separated, then the figures in our tables should follow. Of course, the problem is that we don’t know these probabilities, but if the model is right it ought to be possible to work backwards from the figures observed to a set of probabilities and a formula, and then forward again to predict these figures (and any future figures). This may sound a trivial task, but it is not; and if our assumptions are wrong, it will be simply impossible.

One difficulty in doing this analysis is that we depend on the figures supplied by Lascaratou, who makes fewer distinctions than we need. In particular, she does not (and probably could not) classify her clauses according to their focus properties, so we have to make the best of what we do know. The main fact we now have is that preverbal O’s, and only these O’s, are focussed. This allows us to work out the percentage of O’s that are focussed - about 16%. (The figure is approximate because we don’t know how many of the 19 SOV clauses had focussed O; I have assumed an equal split between S and O.) If we now assume that each element has the same chance of being focussed - a really wild assumption, of course, but a reasonable one - then we can calculate how many of Lascaratou’s clauses with each VO order had focussed S, focussed V or no focus at all. The result of these calculations is shown in Table 6.
<table>
<thead>
<tr>
<th></th>
<th>observed total</th>
<th>guessed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>total</td>
</tr>
<tr>
<td>SVO</td>
<td>1246</td>
<td>1255</td>
</tr>
<tr>
<td>VSO</td>
<td>27</td>
<td>31</td>
</tr>
<tr>
<td>VOS</td>
<td>19</td>
<td>31</td>
</tr>
<tr>
<td>OVS</td>
<td>219</td>
<td>239</td>
</tr>
<tr>
<td>OSV</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>SOV</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>total</td>
<td>1540</td>
<td>1577</td>
</tr>
</tbody>
</table>

Table 6: Guessed distribution of focus among the 6 orders.

The figures in this table show the effects of assuming that each element has a 16% chance of being focussed. The other consideration which we are assuming to be relevant is the separation of S and O, which we found in 95% of observed cases. The figures in the table are based on a slightly different figure, of 97%, which produces a better fit with the observed data. When we combine these two assumptions (16% for focussing, 97% for separation), we get the figures shown in the individual cells; for example, 239 under ‘V focussed’ is 16% of 97% of 1540. Adding the figures for each guessed line gives the figures in the ‘guessed total’ column, which are statistically indistinguishable from the observed totals (Chisquare = 4.910, p > .10). In other words, the formula works well, and our analysis survives.

To show that languages don’t have to work like this, let’s look at some figures on Ancient Greek that I have received from Ann Taylor, of the University of Pennsylvania. She tells me that in a sample of finite clauses from Homer, Herodotus and the Gospel according to Luke the orders are distributed as in Table 7.
<table>
<thead>
<tr>
<th></th>
<th>Homer</th>
<th>Herodotus</th>
<th>Luke</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOV, OSV</td>
<td>44%</td>
<td>26</td>
<td>7</td>
</tr>
<tr>
<td>SVO</td>
<td>33</td>
<td>46</td>
<td>60</td>
</tr>
<tr>
<td>OVS</td>
<td>14</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>VSO, VOS</td>
<td>9</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>total words</td>
<td>93</td>
<td>129</td>
<td>94</td>
</tr>
</tbody>
</table>

Table 7: Distribution of orders in Ancient Greek.

To judge by these figures Greek has changed radically as far as the principles underlying word order are concerned, even though it has always been a 'free word-order' language. Homer's principles favoured verb-final orders, which gave way to verb-medial and some verb-initial orders by Luke, but there is no evidence of the modern principle of separation which favours OVS.

The discussion so far has established that it is possible to explain the distribution of clause-orders in Lascaratou's data by assuming just two variables, separation and focussing, each of whose alternative values has a fixed numerical weight. Focussing is clearly a pragmatic matter, but it is equally clear that the way it is expressed is controlled by grammatical rules, because otherwise it is hard to explain why focussed objects have to precede the verb whereas focussed subjects can be anywhere. The rule of separation is another clear example of a grammatical rule. The outcome of the statistical analysis, therefore, is a definition of what we should expect of a grammatical analysis of Modern Greek. It must contain some kind of mechanism for locating focussed objects before the verb (though not necessarily in the clause-initial position assumed by Lascaratou), and another one for the relation which keeps S and O separated. One of the attractions of a statistical approach is that it allows one to approach grammar-writing without preconceptions, which is exactly the opposite of the approach most of us take. Instead of starting with some theoretical tool-kit, and asking how we can apply these tools to the data, we start with the data and work out what kinds of tools are needed.

At this point you might expect an advertisement for my own pet theory, Word Grammar (Hudson 1990). I should very much like to be able to say that the two
mechanisms that we need are already in place in standard versions of Word Grammar, but in all honesty I can’t. This isn’t surprising, because the theory has grown out of data which are mostly from English, and English is clearly different from Greek. I comfort myself with the thought that no other theory seems to be in a better position. I can think of a few WG tricks that would be worth trying out, but I have nothing concrete to report at the moment.

4. Conclusion

The main point I have tried to make is that statistical analysis of texts is a method which linguists should use more than they do. I used the case of word-class frequencies to illustrate the unexpected facts that can be discovered in this way - and, rather obviously, in no other way than this. The discussion of word-order in Greek, on the other hand, showed how it can help in the analysis of rather well-known facts, by providing a criterion against which alternative theories can be tested.

I should like to end by emphasising that I do not see statistical analysis as anything but a tool to complement the more familiar methods. A set of figures is not an explanation; it is just a tool which helps us to know more clearly what it is that needs explaining. For instance, if my interpretation of the word-order data is correct, it establishes the existence of two independent influences decide the order of elements in a Greek clause, and it shows that one influences the position of O in relation to V, while the other influences the position of both S and O in relation to V. But that is as far as these statistics take us. They do not tell us what the sources of these influences are. Similarly, the statistics on word-classes show us that each of the major word-classes is controlled by some set of influences which apply at a remarkably regular rate, but they don’t in themselves identify the influences. The situation is familiar to natural scientists and social scientists, all of whom use statistical data as standard tools to help in theory-development; perhaps it is time for theoretical linguists to follow suit.

References


