Close to final draft of paper that appeared in *Theoretical Linguistics* 28: 151-157.

# Case Distinctions, Rich Verb Agreement, and Word Order Type

# (Comments on Hawkins' paper)[1]

Matthew S. Dryer

University at Buffalo

Hawkins discusses a variety of asymmetries in the area of word order and proposes a variety of explanations in terms of parsing or sentence processing. I am sympathetic overall to the kinds of explanations he proposes (cf. Dryer 1980, 1992). In some instances, however, the data supporting the generalizations he puts forward is rather limited. Because I have a typological database (Dryer 1989, 1991, 1992) containing a large amount of relevant data and because of space limitations, I will concentrate my discussion on empirical evidence from my database bearing on two of Hawkins' empirical claims. The first of these is his claim that the presence of a case distinction between subjects and objects is most frequent in SOV languages, next most frequent in SVO languages, and least frequent is V-initial languages. The second is his claim that the presence of rich verb agreement (agreement with both arguments in a transitive clause) is most frequent in V-initial languages and least frequent in SOV languages.

## 1. Case distinctions between subject and object

Hawkins predicts that SOV languages will distinguish subjects and object by case marking more often than SVO languages and that SVO languages will do so more often than V-initial languages. He cites the data in Table 1 based on Nichols (1986) supporting this, where the

[1] The data in this paper comes from research that was supported by Social Sciences and Humanities Research Council of Canada Research Grants 410-810949, 410-830354, and 410-850540, and by National Science Foundation Research Grant BNS-9011190.

percentages represent the percentages of languages with “case affixes that distinguish at least two arguments in a clause, e.g. nominative and accusative, or absolutive and ergative” (p. 37).

| SOV | SVO | V-initial |
|---|---|---|
| 89% (25/28) | 60% (3/5) | 38% (5/13) |

Table 1

Hawkins’ data for proportions of languages with case distinction

There are two problems here. First, the numbers here are very small: the conclusion is based on a total of only 5 SVO languages. Second, the sample used, Nichols (1986), is not a representative sample (cf. Dryer 1989) and large parts of the world are not represented by it.

My own typological database (Dryer 1989, 1991, 1992) contains data on this for a much larger and more representative sample of languages; it contains data on the relevant typological features for 502 languages. The data comparable to Hawkins’ data in Table 1 is given in Table 2, both in terms of numbers of genera (genetic groups comparable to the subfamilies of Indo-European; cf. Dryer 1989) and in terms of numbers of languages[2]; the percentages represent

[2] As argued in Dryer (1989), the figures based on genera probably give a more reliable picture; however I cite data in both forms to illustrate that the claims made here are supported by other set of figures. As discussed in Dryer (1989), the numbers of genera are really numbers of subgenera in the sense that a given genus will be counted more than once if it contains two or more of the types under discussion. The data in Table 2 includes languages where the case marking distinguishing subject and object involves adpositions as well as case affixes. The spirit of the theoretical discussion by Hawkins predicts that it should not matter whether the case marking takes the form of adpositions or case affixes.

languages where there is some form of case marking, either affixal or adpositional, distinguishing the two arguments in a transitive clause, when these arguments are nouns. The figures based on numbers of genera represent more accurate estimates, since they control for the bias created by multiple languages from the same genus (Dryer 1989); however, I include them here so that the reader can see that my conclusions obtain, whether one counts genera or languages.

| | SOV | SVO | V-initial |
|---|---|---|---|
| % of genera | 62% (85/138) | 20% (16/82) | 41% (15/37) |
| % of languages | 72% (181/253) | 14% (26/190) | 47% (28/59) |

Table 2

Proportions of languages with case distinction in Dryer database

The data in Table 2 show that Hawkins' prediction that SOV languages will have case marking most often is borne out, both in terms of numbers of genera and in terms of numbers of languages; the percentages are considerably higher for SOV languages than they are for either SVO or V-initial languages. On the other hand, his prediction regarding the order of SVO and V-initial languages is not borne out. Regardless of whether one counts genera or languages, the data in Table 2 shows that V-initial languages are much more likely to have case marking on the subject and/or object than SVO languages: the proportion is much higher for V-initial languages, 41% vs. 20% in terms of proportions of genera, 47% vs. 14% in terms of proportions of languages.

Hawkins' prediction that V-initial languages should have case marking least often is based on the fact that the early placement of the verb will minimize the frequency of misassignments as to what is subject and what is object. However, there is a plausible additional factor that probably explains the fact that SVO languages have case marking on subject and/or object least often. As Hawkins notes, transitive clauses with a lexical subject and a lexical object are relatively infrequent.

In addition, as shown by Gilligan (1987), most languages do not require independent pronouns in subject position; the same is probably true, though to a lesser extent, for objects. As a result, the most frequent transitive clauses (apart from ones where both arguments are pronominal) will be ones with the verb and one noun phrase, either subject or object. In an SVO language, it will be easy to tell whether that one noun phrase is subject or object: if it is subject, the clause will take the form NP-V; if it is object, it will take the form V-NP. However, in a V-initial language, it will take the form V-NP, regardless of whether the one noun phrase is subject or object. Without case marking, this structure will be potentially ambiguous. But case marking will disambiguate the structure. SOV languages will be analogous to V-initial languages in this respect: a clause of the form NP-V will be potentially ambiguous if there is no case marking. The use of word order to distinguish subjects and objects predicts that SVO languages will have case marking less often than either SOV or V-initial languages. However, the principle that Hawkins appeals to is a plausible hypothesis for why SOV languages have case marking more frequently than V-initial languages.

## 2. Rich verb agreement

Hawkins also predicts that rich verb agreement (agreement for both subject and object) will be most frequent in V-initial languages and least frequent in SOV languages. He cites the data in Table 3 to support this, where the percentages reflect percentages of languages with rich agreement in data that again comes from Nichols (1986).

| V-initial | SVO | SOV |
|---|---|---|
| 77% (10/13) | 60% (3/5) | 46% (13/28) |

Table 3

Hawkins' data for proportions of languages with rich verb agreement

As before, the numbers cited by Hawkins are small, especially for SVO languages, and Nichols' sample is not representative. In Table 4, I give the data from my database, based on 557 languages.

| | V-initial | SVO | SOV |
|---|---|---|---|
| % of genera | 62% (29/47) | 47% (49/104) | 54% (80/147) |
| % of languages | 56% (48/86) | 44% (94/213) | 49% (140/283) |

Table 4

Proportions of languages with rich verb agreement in Dryer database

The data in Table 4 support Hawkins' prediction that rich agreement should be most frequent in V-initial languages, which exhibit higher percentages, both in terms of genera and in terms of languages. They do not, however, support his prediction regarding SVO languages: the data in Table 4 shows SOV languages exhibiting a higher percentage with rich agreement than SVO languages.

The overall differences in Table 4 are relatively small and it is not clear that they might not simply reflect random variation. As discussed in Dryer (1989), just using totals of genera is not reliable, because there might be a large concentration of one type in one or two areas of the world, and in order to determine whether a more frequent type represents a real linguistic preference, one should ideally find the preference manifesting itself in all parts of the world. In Table 5, I give a breakdown of the numbers of genera in six large continental areas of the world, enclosing the more frequent type within each area in a box.[3] I use "RichAGr" to denote languages with rich

[3] Southeast Asia & Oceania (SEAsia&Oc) includes Sino-Tibetan, Tai-Kadai, Mon-Khmer, Hmong-Mienic, and Austronesian languages; Eurasia includes all of Europe and Asia excluding the

agreement, i.e. languages with agreement for both subjects and object, while ‘NotRichAgr’ represents languages that do not have agreement for both subjects and objects (though possibly agreement with one of these); “V-1” stands for V-initial languages.

above groups. The other abbreviations are Aus-NewGui (Australia and New Guinea), NAmer (North America, including Mexico, and Guatemala), and SAmer (South America, plus Central America other than Guatemala). If a genus contains languages of more than one type, they will be counted in more than one cell in Table 5. The proportions of genera are thus strictly speaking proportions of “sub-genera”, where subgenera of a genus are defined as subsets of the languages in a genus that differ from each other with respect to the types under examination, but the same within each subgenus.

| | Africa | Eurasia | SEAsia&Oc | Aus-NewGui | NAmer | SAmer | Total |
|---|---|---|---|---|---|---|---|
| V-1&RichAgr | 5 | 1 | 4 | 1 | $\boxed{15}$ | $\boxed{3}$ | 29 |
| V-1&NotRichAgr | $\boxed{7}$ | 1 | 4 | $\boxed{2}$ | 3 | 1 | 18 |
| Proportion RichAgr | .42 | .50 | .50 | .33 | .83 | .75 | Avg=.56 |

| | Africa | Eurasia | SEAsia&Oc | Aus-NewGui | NAmer | SAmer | Total |
|---|---|---|---|---|---|---|---|
| SVO&RichAgr | 21 | 2 | 5 | $\boxed{8}$ | $\boxed{7}$ | $\boxed{6}$ | 49 |
| SVO&NotRichAgr | $\boxed{24}$ | $\boxed{6}$ | $\boxed{16}$ | 5 | 1 | 3 | 55 |
| Proportion RichAgr | .47 | .25 | .24 | .62 | .87 | .67 | Avg=.52 |

| | Africa | Eurasia | SEAsia&Oc | Aus-NewGui | NAmer | SAmer | Total |
|---|---|---|---|---|---|---|---|
| SOV&RichAgr | 8 | 11 | 3 | $\boxed{29}$ | $\boxed{19}$ | $\boxed{10}$ | 80 |
| SOV&NotRichAgr | $\boxed{13}$ | $\boxed{18}$ | $\boxed{8}$ | 15 | 6 | 7 | 67 |
| Proportion RichAgr | .38 | .38 | .27 | .66 | .76 | .59 | Avg.=.51 |

Table 5

Rich agreement and word order type: numbers of genera by area

Let us focus on the hypothesis that V-initial languages are more likely to have rich agreement than SOV languages. Let us do this by determining how many areas exhibit a stronger preference for rich agreement in V-initial languages compared to SOV languages. We can do this by comparing the proportions of genera with rich agreement in each of the six areas. Table 6 extracts the data from the third and ninth lines of Table 5, and encloses the higher proportion for each area in a box.

| | Africa | Eurasia | SEAsia&Oc | Aus-NewGui | NAmer | SAmer | Total |
|---|---|---|---|---|---|---|---|
| V-1 | $\boxed{.42}$ | $\boxed{.50}$ | $\boxed{.50}$ | .33 | $\boxed{.83}$ | $\boxed{.75}$ | Avg=.56 |
| SOV | .38 | .38 | .27 | $\boxed{.66}$ | .76 | .59 | Avg.=.51 |

Table 6

Proportions of genera exhibiting rich agreement

Table 6 shows a higher proportion of rich agreement in five of the six areas. Furthermore, the one area exhibiting the opposite trend, Australia-New Guinea, has only three genera containing V-initial languages so the proportion for this area is based on a small number of genera. Since we find the preference manifested in five of the six areas, and since the one area that exhibits the opposite tendency involves a rather small number of V-initial languages, we can tentatively conclude that Hawkins' prediction that rich agreement should be higher among V-initial languages than among SOV languages is supported.

Let us now do a similar comparison of SVO and SOV languages, to see how many areas exhibit a stronger preference for rich agreement in SOV languages. Again, we can do this by comparing the proportions of genera with rich agreement in each of the six areas. Table 7 extracts the data for this from Table 8, and encloses the higher proportion in a box.

| | Africa | Eurasia | SEAsia&Oc | Aus-NewGui | NAmer | SAmer | Total |
|---|---|---|---|---|---|---|---|
| SVO | $\boxed{.47}$ | .25 | .24 | .62 | $\boxed{.87}$ | $\boxed{.67}$ | Avg=.52 |
| SOV | .38 | $\boxed{.38}$ | $\boxed{.27}$ | $\boxed{.66}$ | .76 | .59 | Avg.=.51 |

Table 7

Proportions of genera exhibiting rich agreement

Table 7 shows that there are three areas in which the proportion is higher among SVO languages, and three areas in which the proportion is higher among SOV languages. We can conclude from this that there is no reason to believe that the difference in Table 4 in terms of total numbers of genera reflects any linguistic preference for rich agreement among SOV languages. The higher numbers in Table 4 turn out to reflect nothing more than the fact that SOV languages happen to be considerably more common than SVO languages in the areas in which rich verb agreement is more common (Australia-New Guinea, North America, and South America).

One thing that is worth drawing attention to is that Table 5 shows an interesting parallelism between SVO and SOV languages. For both word order types, languages with rich verb agreement are more common in Australia-New Guinea, North America, and South America, while languages lacking rich verb agreement are more common in Africa, Eurasia, and Southeast Asia & Oceania. For both word order types, the proportion of genera with rich agreement is highest in North America than it is for any of the other five areas (.87 for SVO languages, .76 for SOV languages). This corresponds to the well-known fact that polysynthetic languages are especially common in North America. Similarly, for both word order types, the proportion with rich agreement is lowest in Southeast Asia & Oceania than it is for the other five areas. This similarly corresponds to the well-known fact that many languages of Southeast Asia are isolating. The general moral of these observations is that richness of agreement seems to correlate more strongly with linguistic area than with word order type.

The data thus do not provide a basis for ranking SVO languages relative to SOV languages. It also turns out that is there is not a significant difference between V-initial languages and SVO languages either, as the reader can confirm by comparing the proportions in Table 5: the proportion is higher for V-initial languages in three areas and for SVO languages in three areas. The data thus fail to support Hawkins' prediction that SVO languages should be intermediate between V-initial and SOV languages in terms of frequency of rich agreement, but they do not provide evidence against it either.[4]

## References

Dryer, Matthew S. 1980. The positional tendencies of sentential noun phrases in universal grammar. *Canadian Journal of Linguistics* 25: 123-195.

[4] My conclusions here differ from those of another study on the relationship between agreement and word order type, namely Siewierska and Bakker (1996), based on a sample of 237 languages. While it would not be appropriate to discuss their study at length here, I should note that their results differ from mine in that they do not find a difference between V-initial and SOV languages with respect to the frequency of rich agreement and they do find both of these to have rich agreement more often than SVO languages. The source of these different results warrants further investigation.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13: 257-292.

Dryer, Matthew S. 1991. SVO languages and the OV:VO typology. *Journal of Linguistics* 27: 443-482.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68: 81-138.

Gilligan, Gary. 1987. *A Cross-Linguistic Approach to the Pro-Drop Parameter*. Unpublished University of Southern California dissertation.

Nichols, Johanna. 1986. Head marking and dependent marking grammar. *Language* 62: 56-119.

Siewierska, Anna, and Dik Bakker. 1996. The distribution of subject and object agreement and word order type. *Studies in Language* 20: 115-161.