# Clause relationships in spoken and written English

Sidney Greenbaum and Gerald Nelson
*Survey of English Usage*
*University College London*

There are opposing views on whether speech or writing is more complex syntactically. We investigated the complexity of clause relationships in a range of spoken and written text categories: spontaneous conversations, broadcast discussions, unscripted monologues, personal handwritten letters, academic writing, and non-academic writing. Conversations proved to be the most distinctive category. It had the highest percentage of simple clauses and the lowest percentage of both subordination and coordination. For all the other categories there is not a sharp distinction between speech and writing in any of the measures that were applied.

In Molière's *Le Bourgeois Gentilhomme*, Monsieur Jourdain is astounded to hear that he has been talking prose all his life. As corpus linguists we are equally incredulous.

We know that speech and writing differ in the opportunities they allow for reconsideration. Written communication is essentially spatial: it typically presents an edited version that has been subjected to one or more revisions. Spoken communication is essentially temporal: it does not allow deletions, only on-the-spot corrections and afterthoughts. We would expect these differences to have consequences for the syntax of speech and

*Functions of Language* 2:1 (1995), 1–21. DOI 10.1075/fol.2.1.02gre
ISSN 0929–998X / E-ISSN 1569–9765 

writing. We might also expect that the constraints imposed by the limitations of short-term memory in the processing of speech would have an effect on spoken syntax, so that (for example) the syntax of clause relationships will be simpler in speech.

There is indeed a widespread view that the spoken language is characterized by parataxis (coordination and juxtaposition of clauses) and the written language by hypotaxis (subordination of clauses). For example, Wallace Chafe (1986) maintains that in general the spoken language is 'fragmented', 'typically constructed of relatively independent clauses and clause fragments', and that 'often the clauses are linked by "and"' (28); in contrast, according to Chafe, writers have the means to 'create elaborate, integrated sentences of a sort that is rare in speech' (29). Chafe concludes: 'Whereas speakers perforce construct sentences on the run, writers can linger over them, fashioning them into objects of a complexity that can sometimes be overwhelming' (29).

Opposing views are propounded by Halliday. He argues that 'speech is no less complex than writing' and that writing has 'rather simple grammatical frames', and adduces an example from the spoken language where 'the sentence structure is highly complex, reaching degrees of complexity that are rarely attained in writing' (Halliday 1994: xxiv). Halliday maintains that writing has a different kind of complexity. It gains its complexity through its vocabulary, 'the packing together of lexical content' in what he calls 'rather simple grammatical frames'. Halliday is convinced that 'it is only in spoken language, and specifically in natural, spontaneous interaction, that the full semantic (and therefore grammatical) potential of the system is brought into play. If you listen grammatically, you will hear sentences of far greater complexity than can ever be found in writing' (Halliday 1992: 62).

Clearly, the opinions expressed by Chafe and Halliday on the relative complexity of the spoken and written language are flatly contradictory. In part their difference may depend on how they calculate complexity (see Note 2).

Halliday's position is supported in a study by Karen Beaman (1984) based on spoken and written narratives. She found a distinctly higher percentage of coordinated sentences (without any subordinate clauses) in writing (38% versus 25%). On the other hand, subordinate sentences were more frequent in speech (18% versus 13%). The spoken narratives also had more coordinate sentences containing subordinate clauses (27% versus 18%). She concludes that 'on the assumption that subordination

implies complexity' her results show that 'spoken narrative is on the whole just as complex as, if not more in some respects, than written narrative' (78).

Beaman's study is restricted to narratives. These do not represent naturally occurring language since they were elicited in an experimental environment in which the subjects — all of them women — were asked to tell what they had just seen in a film. As a result, the narratives did not exhibit the variation that might be expected in samples of natural speech or writing. Leaving aside these reservations and possible effects of a particular method of calculating complexity, we wonder whether the distinctions that Beaman found in narratives apply to speech and writing generally.

We have recently started a research project to investigate clause relationships in English, focusing in particular on the spoken language. One of our aims is to contrast the uses of spoken and written English in a variety of registers. We decided to examine the evidence for the opposing views of Chafe and Halliday on clause complexity in speech and writing.

For our research project we have selected a sub-corpus of spoken and written texts drawn from the British component of ICE (the International Corpus of English). We call this the Leverhulme Corpus, because the two-year research project is funded by the Leverhulme Trust.[1]

SPEECH (30)
- Spontaneous conversations (20)
- Unscripted monologues (5)
- Broadcast discussions (5)

WRITING (12)
- Personal letters (4)
- Academic writing (4)
  - humanities (1)
  - social sciences (1)
  - natural sciences (1)
  - technology (1)
- Non-academic writing (4)
  - humanities (1)
  - social sciences (1)
  - natural sciences (1)
  - technology (1)

Table 1: *Composition of the Leverhulme Corpus*

The Leverhulme Corpus consists of 42 texts, each containing about 2,000 words. Most of the texts are from speech recordings: 20 spontaneous conversations, 5 broadcast discussions, and 5 unscripted monologues. The written component consists of 4 texts of handwritten personal letters and 8 printed texts, which are divided equally between academic and non-academic writing. In each set of printed material, four categories were represented: humanities, social sciences, natural sciences, and technology. The texts were selected to provide a variety of speakers and writers according to sex, age, and educational level. Some of the conversations were between equals, and others were between those in an unequal relationship, such as teacher/student and doctor/patient (see Appendix). Table 1 displays the composition of the Leverhulme Corpus.

We annotated the Leverhulme Corpus manually for clause relationships. The units we selected for analysis were based on syntax. For the written component of our corpus we could not, for example, use orthographic sentences, since they may represent rhetorical units rather than syntactic units. For example, an orthographic sentence may consist of two or more syntactic sentences, separated by perhaps semicolons:

(1) In their own estimation their rule rested on right and not on mere force; they were accepting the established doctrine of Greek political philosophy that government exists for the welfare of the governed. [W2A-001-42]

Or an orthographic sentence may consist of just a fragment that belongs to a preceding sentence, but is isolated orthographically to represent an intonation break:

(2) It may be, in addition, that it was necessary for Charles Dickens to keep on working in order to prove that his father was truly an 'insolvent' person. And, as they walked together, around them sprang the morning life of the metropolis; the clerks and office boys already streaming in to the city from the outlying areas, the apprentices sweeping their shops and watering the pavements outside, the children and servants already crowding the bakers' shops, the fast coaches going on their appointed rounds. But, for the young Dickens, above all the sun rising over another blank day, over the dreariness and the subdued low pain of loss.

*A loss made all the greater, from his own account, by the spectacle of seeing his older sister win a prize — or, rather, two prizes.* [W2B-006-50ff]

Less unusually and less surprisingly, coordination of clauses may also cross orthographic sentences, as in (2), where the third sentence consists of a participle clause linked by *but* to a series of participle clauses in the preceding sentence.

Chafe's claim that spoken syntax is much simpler than written syntax is grounded on what he takes to be the units for analysis in speech and writing. His speech analysis is based — at least in part — on intonation units and on closures by what he terms sentence-final intonation, though the principal criterion for his sentence in speech seems to be a unit that expresses a single 'center of interest' (Chafe 1980: 26-38). We consider this approach too subjective since it requires decisions on what are the centres of interest in a discourse.[2]

It is appropriate to begin with a description of our methodology. We base our analysis on the clause, which can be identified as consisting of well-established relational elements such as subject, verb, and verb complements. Clauses may be finite, non-finite, or verbless (cf. Quirk et al. 1985: 14.5-9). Our fundamental units are the simple clause and the complex clause. A simple clause contains no subordination, a complex clause contains one or more subordinate clauses.[3] For example, (3) is a simple unit and (4) a complex unit:

(3) The cellular anatomy of the peripheral nervous system renders it vulnerable to injury. [W2A-026-2]
(4) Living in the Gulf has meant living with oil. [W2B-029-13]

There is only one clause in the simple unit (3); it is not subordinated to any other clause, nor is another clause subordinated to it. In the complex unit (4) both the subject *living in the Gulf* and the object *living with oil* are subordinate clauses, in both instances participle clauses.

Clause units may be linked to each other within a clause cluster. Minimally, a clause cluster is realized by one clause unit (either a simple unit or a complex unit). Hence, (3) and (4) are clause clusters. They differ in that (3) is a simplex cluster and (4) is a complex cluster. But a clause cluster may consist of a combination of units linked by coordination, in which case it is a compound cluster. The clusters in (5) and (6) are compound clusters:

(5) Commercial telephone ringers usually employ a piezo-electric device to create sound *but* in this application this proved somewhat feeble in both volume and resonant quality [W2B-032-50]

(6) These are the sort of things that are slaughtered <,> in honour of the gods <,> *and* you obviously don't do that inside the building [S2A-024-95f.]

The cluster in (5) involves a coordination by *but* of a complex unit (containing the infinitive purpose clause *to create sound*) with a simple unit. In (6) the cluster again consists of a combination of a complex unit and a simple unit, but this time subordination takes the form of a restrictive relative clause *that are slaughtered in honour of the gods*.[4]

Clause clusters may be identified with canonical sentences, and they have proved to be essential segments in our calculations of coordination and subordination. We have therefore taken care to be as precise as we can in delimiting them. We have reproduced two extracts from the Leverhulme Corpus, one from academic writing and the other from an unscripted monologue (Tables 2 and 3).

In addition to clause units we recognize paratactic clauses, clause fragments, non-clauses and incomplete units. Paratactic clauses (in this restricted sense) are those that are adjacent to or inserted in other structures but are neither coordinated with them nor subordinated to them. The paratactic clauses are either tag questions or parenthetics. The most frequent are the discourse markers *I mean, you know, you see*.[5]

Clause fragments are usually noun phrases or prepositional phrases that serve as responses to a previous clause. Their clause functions are recognizable if they are analysed as elliptical clauses, in which case the ellipted parts are recoverable from the preceding context. Examples of clause fragments are given in the conversation extracts (7)-(9):

(7) B: What else did Linda have to say for herself <,>
A: *Oh a lot* [S1A-010-191f.]

(8) A: What time does she normally have a lesson
B: *Three*
*Three o'clock* [S1A-083-44ff.]

(9) B: You know well certain certain parts of it that other people wouldn't normally understand
A: Oh I see
B: *Sort of like jargons slangs*
A: *Sort of* [S1A-015-168ff.]

| | | Unit type | Cluster type |
|---|---|---|---|
| 1. | In NDT the conventional method of presenting 3D data is as 3 orthogonal views, one view for each of the x, y, and z directions (B, C and D-Scans). | complex clause | complex |
| 2a. | Ideally these representations would show a perfectly formed image of the defect | simple clause | compound |
| 2b. | but ultrasonics does not afford this luxury | simple clause | |
| 2c. | and the B, C, and D-Scans typically illustrate only an abstract view of the defect. | simple clause | |
| 3. | For example, a smooth linear defect angled at 45º to the surface and scanned at 0º would be displayed as two diffraction arcs on the B-Scan as Figure 1b illustrates. | complex clause | complex |
| 4. | Since the same B-Scan pattern would have been achieved if there were two point defects at the position of the defect tips, this illustrates that data from a single probe is insufficient for defect classification. | complex clause | complex |
| 5a. | Reference to the B, C, and D-Scans for other probes is usually necessary to deduce the nature of the defect, | complex clause | complex |
| 5b. | for example, reference to the B-Scan for the 45º probe ( Figure 1c ) would allow more accurate categorisation. | simple clause | simplex |

Table 2: *Extract from academic writing* [W2A-036-2ff]

| | | Unit type | Cluster type |
|---|---|---|---|
| 1 | The study of these prognostic factors plays an important role in an analysis of clinical studies <,,> | simple clause | simplex |
| 2a. | The whole purpose of a clinical trial is to advance knowledge | complex clause | compound |
| 2b. | and therefore it is important that all randomised clinical trials should be published irrespective of their results | complex clause | |
| 3. | Unfortunately there tends to be a bias towards publishing only positive results | complex clause | complex |
| 4. | There is also a danger of publishing trials with too small patient numbers which can produce false positive results and exaggerated treatment claims | complex clause | complex |
| 5. | Clinicians need to make critical evaluations of the papers in the medical journals <,,> | simple clause | simplex |
| 6. | The separation of patients into good and poor prognostic groups can assist in the design of future studies either as radical curative studies or studies of palliative intent only | simple clause | simplex |
| 7a. | Unfortunately in cancer research small improvements between treatments is all that can realistically be expected | complex clause | compound |
| 7b. | and therefore studies with large patient numbers are needed often in excess of four hundred patients | simple clause | |

Table 3: *Extract from an unscripted monologue* [S2A-033-27ff.]

Non-clauses are independent units that do not have the structure of clauses. In printed texts, they are usually headings; in letters, they are usually addresses, dates, salutations, sign-offs; in spoken texts, they may be voiced hesitations such as *uh* and *uhm*, interjections such as *oh* and *ah*, and formulaic expressions and reaction signals such as *hello*, *yes*, *no*. A non-clause may consist of more than one word, as in addresses on letters, or combinations such as *ah yes* or *oh right*, or a series of voiced hesitations. Non-clauses do not enter into clause coordination. If the following unit begins with a coordinator, we regard it as the start of a new cluster. This point is illustrated in (10), which is not analysed as a coordination of the non-clause *yes* and a simple unit. We count the clause beginning with *but* as a simplex cluster, since we are investigating complexity of clause structures.

(10) A: You're back this week aren't you
B: *Yes but I'm off Thursday and Friday* [S1A-061-7f.]

Syntactically incomplete clauses only appeared in the spoken texts within the Leverhulme Corpus. Included in this category are only those incomplete clauses whose status as simple or complex unit is impossible to determine.[6] The speaker in (11) has not completed two of his clauses, and their status is indeterminate:

(11) This is what uh we would call a picture <,> or you would call a picture <,> *and it's actually* <,> *eventually it will be* <,>
Oh no it won't [S2A-029-71f.]

We have applied chi-square tests to the raw data in the Tables that follow, and we have regarded the distinctions as significant whenever the level is less than 0.001.

Table 4 displays the numbers of simple units and complex units in the Leverhulme Corpus whether they constitute independent clusters or parts of compound clusters. In addition, the Table separately lists the numbers of paratactic clauses, clause fragments, non-clauses, and incomplete clauses.

As we might expect, Table 4 shows that paratactic clauses, fragments, non-clauses, and incomplete clauses predominantly occur in conversations. Together the four types amount to 50.4% of units in conversation, yielding a significantly higher number than for any other category. If we exclude these, we obtain a clearer view of the relative

| Category | Simple clauses | Complex clauses | Para-tactic clauses | Frag-ments | Non-clauses | Incom-plete units | Total |
|---|---|---|---|---|---|---|---|
| Non-academic writing | 36.9% (176) | 58.7% (280) | 1.0% (5) | 0.5% (2) | 2.9% (14) | - (0) | 100% (477) |
| Academic writing | 40.8% (210) | 50.9% (262) | 0.4% (2) | 1.3% (7) | 6.6% (34) | - (0) | 100% (515) |
| Letters | 39.4% (299) | 40.1% (304) | 1.8% (13) | 5.0% (38) | 13.7% (104) | - (0) | 100% (758) |
| WRITTEN | 39.1% (685) | 48.3% (846) | 1.2% (20) | 2.7% (47) | 8.7% (152) | - (0) | 100% (1,750) |
| Monologues | 30.7% (225) | 49.5% (363) | 4.4% (32) | 1.5% (11) | 12.7% (93) | 1.2% (9) | 100% (733) |
| Broadcast discussions | 29.0% (263) | 39.9% (362) | 7.3% (66) | 2.2% (20) | 21.1% (192) | 0.5% (5) | 100% (908) |
| Conversations | 31.3% (2,113) | 18.3% (1,233) | 9.4% (634) | 7.0% (469) | 31.2% (2,100) | 2.8% (192) | 100% (6,741) |
| SPOKEN | 31.0% (2,601) | 23.3% (1,958) | 8.7% (732) | 6.0% (500) | 28.5% (2,385) | 2.5% (206) | 100% (8,382) |
| Total | 32.4% (3,286) | 27.7% (2,804) | 7.4% (752) | 5.4% (547) | 25% (2,537) | 2.1% (206) | 100% (10,132) |

Table 4: *Units*

frequencies of simple and complex clauses. Table 5 is extracted from Table 4 to demonstrate these relative frequencies.

Table 5 shows that the majority of units in the written data are complex clauses, and the majority in the spoken data are simple clauses.[7] The highest percentage of simple clauses is found in conversations. However, the distinction is not simply between speech and writing, which do not differ significantly. For example, there is no significant difference between conversations and personal letters, or between spoken mono-

logues and non-academic writing. Hence, factors other than the spoken/written dichotomy affect the choice of simple over complex clauses. Clearly, personal letters are closest to conversations in their spontaneity: both text categories involve less planning than other categories of speech or writing. Similarly, personal letters are closest to conversations in their casualness: the two categories relate to private rather than public communication and perhaps exhibit less concern with form. Both degree of planning and level of formality may be factors influencing syntactic complexity.

| **Category** | **Simple clauses** | **Complex clauses** | **Total** |
|---|---|---|---|
| Non-academic writing | 38.6% (176) | 61.4% (280) | 100% (456) |
| Academic writing | 44.5% (210) | 55.5% (262) | 100% (472) |
| Letters | 49.6% (299) | 50.4% (304) | 100% (603) |
| WRITTEN | 44.7% (685) | 55.3% (846) | 100% (1,531) |
| Monologues | 38.3% (225) | 61.7% (363) | 100% (588) |
| Broadcast discussions | 42.1% (263) | 57.9% (362) | 100% (625) |
| Conversations | 63.2% (2,113) | 36.8% (1,233) | 100% (3,346) |
| SPOKEN | 57.0% (2,601) | 43.0% (1,958) | 100% (4,559) |
| Total | 53.9% (3,286) | 46.1% (2,804) | 100% (6,090) |

Table 5: *Simple clauses and complex clauses*

Table 6 takes into account the coordination of clause units.[8] Three types of clause clusters are distinguished: simplexes, complexes, and compounds. The relative frequency of compounds is virtually identical in both modes, but almost half the clusters in the spoken data are simplexes, whereas by far the highest proportion of clusters in the written data are

complexes. However, an examination of text categories within each mode reveals that the broad distinctions do not apply to every category. In the spoken texts there is a wide percentage range both for simplexes (24.9% to 55%) and for compounds (18.1% to 35.5%). In particular, conversations are significantly different from the other categories: they stand out in having the highest proportion of simplexes and the lowest proportions of complexes and compounds.

| **Category** | **Simplexes** | **Complexes** | **Compounds** | **Total** |
|---|---|---|---|---|
| Non-academic writing | 24.7% (86) | 47.4% (165) | 27.9% (97) | 100% (348) |
| Academic writing | 29.7% (118) | 51.4% (204) | 18.9% (75) | 100% (397) |
| Letters | 38.6% (194) | 42.1% (212) | 19.3% (97) | 100% (503) |
| WRITTEN | 31.9% (398) | 46.5% (581) | 21.6% (269) | 100% (1,248) |
| Monologues | 24.9% (96) | 39.6% (153) | 35.5% (137) | 100% (386) |
| Broadcast discussions | 33.1% (146) | 37.9% (167) | 29.0% (128) | 100% (441) |
| Conversations | 55.0% (1,509) | 26.9% (737) | 18.1% (496) | 100% (2,742) |
| SPOKEN | 49.1% (1,751) | 29.6% (1,057) | 21.3% (761) | 100% (3,569) |
| Total | 44.6% (2,149) | 34.0% (1,638) | 21.4% (1,030) | 100% (4,817) |

Table 6: *Clusters*

Table 7 looks more closely at the compound clusters. It differentiates compounds where only simple units are coordinated from compounds where at least one unit contains some subordination. For both spoken and written material, over 70% of the compounds include subordination. Again the speech/writing dichotomy is not the decisive factor, since conversations and academic writing share the distinction of having the

least number of compounds with subordination, whereas monologues and broadcast discussions have the highest number of such compounds.

| **Category** | **Coordinated simplexes** | **Coordination that includes complexes** | **Total** |
|---|---|---|---|
| Non-academic writing | 22.7% (22) | 77.3% (75) | 100% (97) |
| Academic writing | 40.0% (30) | 60.0% (45) | 100% (75) |
| Letters | 25.8% (25) | 74.2% (72) | 100% (97) |
| WRITTEN | 28.6% (77) | 71.4% (192) | 100% (269) |
| Monologues | 10.9% (15) | 89.1% (122) | 100% (137) |
| Broadcast discussions | 18.0% (23) | 82.0% (105) | 100% (128) |
| Conversations | 32.7% (162) | 67.3% (334) | 100% (496) |
| SPOKEN | 26.3% (200) | 73.7% (561) | 100% (761) |
| Total | 26.9% (277) | 73.1% (753) | 100% (1,030) |

Table 7: *Compound Clusters*

In Table 7 we looked at the presence of subordination in just compound clusters. Now we consider the presence of subordination in all clusters where it appears — whether they are compound clusters or complex clusters.

Table 8 brings together data from Tables 6 and 7 to differentiate clusters without subordination from clusters that include subordination. The written component of the corpus has a much higher percentage of clusters with subordination than clusters without, whereas the reverse applies to the spoken component. However, the difference is caused entirely by the conversations, which are significantly different from all

the rest. The other two spoken categories follow the same direction as the written categories. But the written component also shows internal divisions, the category of letters (which resemble conversations in other respects, cf. Table 5) having the highest percentage of clusters without subordination. Monologues are also noteworthy in having a higher percentage of clusters with subordination than any of the written categories.

| **Category** | **Clusters without subordination** | **Clusters with subordination** | **Total** |
|---|---|---|---|
| Non-academic writing | 31.0% (108) | 69.0% (240) | 100% (348) |
| Academic writing | 37.3% (148) | 62.7% (249) | 100% (397) |
| Letters | 43.5% (219) | 56.5% (284) | 100% (503) |
| WRITTEN | 38.1% (475) | 61.9% (773) | 100% (1,248) |
| Monologues | 28.8% (111) | 71.2% (275) | 100% (386) |
| Broadcast discussions | 38.3% (169) | 61.7% (272) | 100% (441) |
| Conversations | 60.9% (1,671) | 39.1% (1,071) | 100% (2,742) |
| SPOKEN | 54.7% (1,951) | 45.3% (1,618) | 100% (3,569) |
| Total | 50.4% (2,426) | 49.6% (2,391) | 100% (4,817) |

Table 8: *Clusters with or without subordination*

So far we have considered whether or not a cluster has subordination. Now we turn to the question of the number of subordinate clauses. Table 9 shows the number of subordinate clauses in each category in relation to the number of clusters in that category. There is a higher subordination ratio in writing than in speech, but within the spoken component there are wild fluctuations, producing the highest and the lowest ratios of all the categories. In view of the results that we had already obtained, we were not

surprised to find a low ratio of subordinate clauses in conversation (an average of 0.7 per cluster), and the highest ratio in monologues and broadcast discussions (averages of 2.0 and 1.9 respectively).

| **Category** | **Number of subordinate clauses** | **Total number of clusters** | **Subordinate clauses per cluster** |
|---|---|---|---|
| Non-academic writing | 568 | 348 | 1.6 |
| Academic writing | 481 | 397 | 1.2 |
| Letters | 571 | 503 | 1.1 |
| WRITTEN | 1,620 | 1,248 | 1.3 |
| Monologues | 790 | 386 | 2.0 |
| Broadcast discussions | 828 | 441 | 1.9 |
| Conversations | 2,054 | 2,742 | 0.7 |
| SPOKEN | 3,672 | 3,569 | 1.0 |
| Total | 5,292 | 4,817 | 1.1 |

Table 9: *Number of subordinate clauses*

We examined the number of coordinated clauses at two levels: clause units (simple or complex) that are coordinated and subordinate clauses that are coordinated. Table 10 shows that the written component had more coordinated clause units than the spoken component, but within both components there was considerable variation. The variation is particularly conspicuous in the spoken material, where we find the highest percentage of coordinated units (in monologues) and the lowest percentage (in conversations). Table 10 also reports on the number of coordinated subordinate clauses. Broadcast discussions have the highest percentage and academic writing has the lowest, but the low frequency of this type of coordination makes detailed comparisons unreliable.

| Category | Number of units coordinated | Total number of units | Number of subordinate clauses coordinated | Total number of subordinate clauses |
|---|---|---|---|---|
| Non-academic writing | 43.6% (208) | 477 | 5.3% (30) | 568 |
| Academic writing | 30.3% (156) | 515 | 5.0% (24) | 481 |
| Letters | 27.8% (211) | 748 | 7.5% (43) | 571 |
| WRITTEN | 32.8% (575) | 1,750 | 6.0% (97) | 1,620 |
| Monologues | 50.3% (369) | 733 | 7.0% (55) | 790 |
| Broadcast discussions | 34.9% (317) | 908 | 10.1% (84) | 828 |
| Conversations | 18.0% (1,216) | 6,741 | 5.3% (110) | 2,054 |
| SPOKEN | 22.7% (1,902) | 8,382 | 6.8% (249) | 3,672 |
| Total | 24.4% (2,477) | 10,132 | 6.5% (346) | 5,292 |

Table 10: *Number of coordinated clauses*

In these preliminary investigations of clause relationships in the Leverhulme Corpus, we have focused on whether or not clauses in a spoken or written discourse are simply juxtaposed or are related by coordination or subordination. We have borne in mind the conflicting views of Chafe and Halliday on the relative complexity of spoken and written English.

Our results do not support a sharp distinction between speech and writing in any of the measures that we have applied. Not only does each mode exhibit considerable internal variation, but also the ranges for each measure that we investigated — with one exception — overlap across the modes. The exception applies to complex clusters (Table 6): they occur

more frequently in the written categories than in the spoken categories. The only consistent general result is that conversations are always distinguished from the other categories. Proportionately, conversations have less coordination and less subordination than any other category. Insofar as conversations are the most typical and the most frequent use of speech, Chafe is correct in his view that there is less complexity in the spoken language than in the written language. However, both monologues and broadcast discussions are always distinct from conversations and they tend to be closer to academic and non-academic writing. Unlike everyday conversations, monologues and broadcast discussions are public in that the speakers are aware of an audience and therefore are more concerned with form. Though broadcast discussions resemble conversations in being interactive, they differ in being controlled by one of the participants. It is clear that factors other than the speech/writing difference affect the use of coordination and subordination in discourse. These might include the degree of planning and the level of formality.

We expect to conduct further studies that may offer additional suggestions for the results we have so far obtained; in particular, we wish to look at individual texts within each category to examine the extent of internal variation and to note correlations with biographical details of writers and speakers. We intend to calculate the relative frequencies of different types of subordinate clauses and their positions within clusters and to take into account the varying levels of subordination. Other objectives that we have in mind are investigations into the discourse functions of clause relationships.

Our results suggest that if Monsieur Jourdain had been speaking English, he would not have been talking prose. On the other hand, although this paper started life as a conference talk, we are confident that in the course of its transformation it has acquired the clause complexity of academic prose.[9]

*Authors' address:*
Sidney Greenbaum and Gerald Nelson · Survey of English Usage · University College London · Gower Street · London WC1E 6BT · UK

## Notes

1. The research for this paper was supported by the grant from the Leverhulme Trust. The ICE project was supported in part by grant R000 23 2077 from the Economic and Social Research Council. We are indebted to Oonagh Sayce for part of the annotation used in this paper.

2. Chafe's principal concern is with the flow of discourse. For that reason, his analysis focuses on idea units, which may or may not correlate with syntactic units and with intonation or punctuation units. Idea units combine to form extended sentences, each of which expresses a single centre of interest. An extended sentence may comprise a sequence of sentences that are not linked by coordination or subordination. For a study of syntactic complexity the extended sentence is too large and difficult to determine.

3. The abbreviated term *complex* (for *complex cluster*) should not be confused with Halliday's *complex* (for *clause complex*). Halliday has a two-term system: (clause) simplex and (clause) complex (cf. Halliday 1992: 344), and a simplex may incorporate embedded clauses (cf. Halliday 1989: 83f.).

4. In the transcriptions of spoken texts, the symbols <,> and <,,> denote short and long pauses respectively. We define a short pause as a perceptible break in phonation which is equivalent in length to a single syllable, uttered at the speaker's tempo. A long pause is any longer break in phonation.

5. Most of the paratactic clauses are discourse markers. The most frequent discourse markers are *I mean*, *you know*, and *you see*; totalling 517, they constitute about 69% of all the paratactic clauses. The discourse markers are in effect fossilized clauses that are being grammaticalized (cf. Mair 1994, 129). Virtually all the paratactic clauses are simple. If they were added to the column of simple clauses in Tables 4 and 5, they would not materially affect the relative percentages of simple and complex clauses.

6. We have not counted repetitions as incomplete clauses. For example, we count *it used to have it used to have a name like the Trocadero* as one simple clause.

7. The ICE tagset characterizes certain combinations as semi-auxiliaries. Some of them are followed by an infinitive, usually with *to*, such as *be about to*, *be going to*, *get to*, *had better*, *seem to*, *start to*. Others are followed by an *-ing* participle; for example: *begin*, *go on*, *keep on*. A major respect in which semi-auxiliaries resemble auxiliaries is that they are semantically independent of the subject (cf. Quirk et al. 1985, 3.29, and 3.45ff.). With respect to the distinction in clause units, the verb that follows a semi-auxiliary is not regarded as the beginning of a new clause, so that *You have to completely suspend belief* [S1A-006-146] — which has the semi-auxiliary *have to* — is a simple sentence. If we had decided on the alternative analysis, merely 149 clauses would have swung from simple to complex (92 of them in conversations) — a switch that would have made little difference to the relative frequencies of simple and complex units in the text

categories.

8. In our investigation of coordination we considered the coordinators *and, but, or* and *nor*. Frequently, however, these have a non-coordinating function, especially in speech, where *and* and *but* often initiate an utterance. For this reason we have not taken them as coordinators if they occur at the beginning of a speaker turn, unless the turn is interrupted by only a very brief utterance, as in the following example:

> A: They've got a thing which is the equivalent of our Aga <,>
> B: Yeah
> A: And they have a conventional cooker as well which they were using <,>
> [S1A-009-179ff.]

In addition, we applied the following criteria: (i) Non-clauses do not enter into coordination. See example (10). (ii) In writing, *and, but, or* and *nor* occurring at the beginning of a paragraph were not counted as coordinators.

9. The article was first presented as a paper at the 15th ICAME Conference, 18-22 May 1994, Aarhus, Denmark.

## References

Beaman, K. (1984) Coordination and subordination revisited: Syntactic complexity in spoken and written discourse. In D. Tannen (ed.), *Coherence in Spoken and Written Discourse*. Norwood, N.J.: Ablex. 45-80.

Chafe, W. (1980) The deployment of consciousness in the production of narrative. In W. Chafe (ed.) *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, N.J.: Ablex. 9-50.

Chafe, W. (1986) Writing in the perspective of speaking. In C. R. Cooper and S. Greenbaum (eds.) *Studying Writing*. Beverley Hills: Sage. 12-39.

Chafe, W. (1992) The importance of corpus linguistics to understanding the nature of language. In J. Svartvik (ed.), 79-97.

Halliday, M. A. K. (1989) *Spoken and Written Language*. Oxford: Oxford University Press.

Halliday, M. A. K. (1992) Language as system and language as instance: The corpus as a theoretical construct. In J. Svartvik (ed.), 61-77.

Halliday, M.A.K. (1994) *An Introduction to Functional Grammar* (2nd edn.) London: Arnold.

Mair, C. (1994) Is *see* becoming a conjunction? The study of grammaticalisation as a meeting ground for corpus linguists and grammatical theory. In U. Fries, G. Tottie, P. Schneider (eds.) *Creating and Using English Language Corpora*. Amsterdam: Rodopi. 127-137.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Svartvik, J. (ed.) (1992) *Directions in Corpus Linguistics.* Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991. Berlin: Mouton de Gruyter.

## Appendix

### Source Texts in the Leverhulme Corpus

**Conversations (1990-1993):**

S1A-005: Two female university students
S1A-006: Male & female colleagues
S1A-009: Mother & son
S1A-010: Mother & daughter, conversation during a game of Scrabble
S1A-013: Conversation among four teachers during a publisher's market research discussion
S1A-015: Male & female friends
S1A-024: University professor & PhD candidate
S1A-028: Family conversation during a birthday party
S1A-031: Two female friends
S1A-033: University Careers Officer and male student
S1A-036: Two female colleagues
S1A-051: Four doctor-patient consultations
S1A-052: Conversation between a photojournalist and his biographer
S1A-059: Consultation with University student counsellor
S1A-061: Two male colleagues' lunchtime conversation
S1A-067: Two female friends
S1A-075: Psychology research interview
S1A-080: Two female friends
S1A-083: Two female tennis coaches
S1A-090: Students' conversations

**Broadcast discussions:**

S1B-024: *Start the Week*, BBC Radio 4, 21-7-91
S1B-026: *Midweek with Libby Purves*, BBC Radio 4, 15-5-91
S1B-027: *Question Time*, BBC 1 TV, 17-1-91
S1B-028: *The Persistence of Faith*, BBC Radio 4, 27-1-91
S1B-035: *Any Questions?*, BBC Radio 4, 9-11-90

**Monologues:**

S2A-024: Patsy Vanags, "Greek Temples", British Museum Public Lecture, 1-5-91
S2A-027: Prof. Hannah Steinberg, "An Academic's Path Through the Media", UCL Lunchtime Lecture, 5-3-91
S2A-029: Three 5-minute presentations by UCL staff members during staff training
S2A-033: Three 5-minute presentations by UCL staff members during staff training
S2A-037: Dr D.M. Roberts, "The Relationship between Industrial Innovation and Academic Research ", UCL Lunchtime Lecture, 15-10-91

**Social letters:**

W1B-006: Six letters between friends
W1B-008: Six letters to female friends
W1B-012: Two letters to male friend
W1B-014: Eight letters between friends

**Academic writing:**

W2A-001: Brunt, P.A., *Roman Imperial Themes*, pp. 110-117 (Oxford: Clarendon Press 1990)
W2A-016: Shannon, John and Chris Howe, "Controlling a Growing Firm", *International Journal of Project Management*, vol. 8 (1990), pp .163-166
W2A-026: Smith, P.J., "Nerve Injury and Repair", in F.D. Burke, D.A. McGrouther and P.J. Smith (eds.), *Principles of Hand Surgery*, pp. 143-153 (London: Longman 1990)
W2A-036: McNab, A. and Iain Dunlop, "AI Techniques Applied to the Classification of Welding Defects from Automated NDT Data", *British Journal of Non-Destructive Testing*, vol. 33 (1991), pp. 11-16

**Non-academic writing:**

W2B-006: Ackroyd, Peter, *Dickens*, pp. 83-88 (London: Sinclair-Stevenson 1990)
W2B-012: Lord Young, *The Enterprise Years: A Businessman in the Cabinet*, pp. 49-55 (London: Headline 1990)
W2B-029: Dipper, Frances, "Earth, air, fire, oil and war", *BBC Wildlife Magazine*, vol.9, (1991), pp. 191-193
W2B-032: Denison, A.C., "Is Anybody There?", *Practical Electronics*, June 1991, pp. 16-20