Applied Linguistics 21/3: 354–375

Measuring Spoken Language: A Unit for All Reasons

[1]PAULINE FOSTER, [2]ALAN TONKYN, AND [3]GILLIAN WIGGLESWORTH

[1]St Mary's University College, [2]University of Reading, [3]Macquarie University

The analysis of spoken language requires a principled way of dividing transcribed data into units in order to assess features such as accuracy and complexity. If such analyses are to be comparable across different studies, there must be agreement on the nature of the unit, and it must be possible to apply this unit reliably to a range of different types of speech data. There are a number of different units in use, the various merits of which have been discussed by Crookes (1990). However, while these have been used to facilitate the analysis of spoken language data, there is presently no comprehensive, accessible definition of any of them, nor are detailed guides available on how to identify such units in data sets. Research reports tend to provide simplistic two-line definitions of units exemplified, if at all, by unproblematic written examples. These are inadequate when applied to transcriptions of complex oral data, which tend not to lend themselves easily to a clear division into units. This paper was motivated by the need each of the three authors felt for a reliable and comprehensively defined unit to assist with the analysis of a variety of recordings of native and non-native speakers of English. We first discuss in very general terms the criteria according to which such a unit might be selected. Next, we examine the main categories of unit which have been adopted previously and provide a justification for the particular type of unit that we have chosen. Focusing on this unit, we identify a number of problems which are associated with the definition and exemplification of units of this type, and give examples of the awkward cases found in actual data. Finally we offer a definition of our unit, the Analysis of Speech Unit (AS-unit), providing adequate detail to address the problematic data analyses we have illustrated.

1. APPROACHES TO SPOKEN LANGUAGE ANALYSIS

Researchers dealing with spoken first or second language data are often seeking to measure the frequency of certain discourse features (such as confirmation checks, clarification requests, self-corrections) or the frequency of certain grammatical features (such as morphemes marking number or case) or else they are seeking to measure quantitatively such dimensions as the relative grammatical accuracy, syntactic complexity and fluency of the language in their data. In order to do this, they need first to segment the data into units against which frequencies and ratios can be calculated. Accordingly, a unit for the segmentation of oral data is an essential tool in

applied linguistics, without which much quantitative analysis of language development would not be possible.

In the assessment of spoken language performance by division into segments, more has often meant better. The 'more' can be seen in two ways: productivity and complexity. In the assessment of young children's or aphasics' first language performance, productivity has been encapsulated in such measures as the Mean Length of Utterance (MLU) devised by Roger Brown (1973), or in the Mean number of Sentences per Turn of the original LARSP[1] procedure (Crystal *et al.* 1976). Complexity in the case of such performances has been measured by a range of indices, including the LARSP structural analysis, the Index of Productive Syntax (Scarborough 1989), and the Type–Token ratio (e.g. Wagner 1985).

Such measures have not gone uncriticized. Snow (1996) has identified a range of problems with the MLU which arise from its focus on product rather than developmental process and its failure to distinguish differences in morphological sophistication among children. Similarly, Richards (1987) has pointed out that although type–token ratios have been widely used to assess children's language development, the measure is unsatisfactory because the ratio inevitably decreases as sample size increases.

In the measurement of older children's first language performance, and in second language measurement, it is difficult to work with the productive dimension alone. The relative cognitive maturity of such performers means that they can be highly productive despite limited resources. They can fashion lengthy continuous turns using a range of communication strategies, such as paraphrases, gestures, and filled pauses. However, such performances are not usually valued as highly as those where high productivity is supported by relatively high complexity, or a wide repertoire.

Other approaches have focused on units or 'chunks' of spoken language within lengthy turns. These can be linked to psycholinguistic processes—more specifically, planning processes. A number of researchers and commentators (e.g. Butterworth 1980; Chafe 1980; Garman 1990) have noted that both macro- and micro-planning processes are involved in composing speech. The former may cover quite long (e.g. multi-'sentence') stretches of speech, while the latter are associated with shorter units similar to the clause or sentence. Regarding the shorter units, from the planning perspective, the more proficient speaker will be the person who can keep track of more complex micro-units. This ability allows the speaker to produce a more complex message in a shorter time span. This may improve information flow and free working memory capacity for other communicative tasks, such as monitoring an interlocutor's reactions, and adjusting the style and/or sequence of information. Keeping track of more complex micro-units will, in turn, involve being aware of the syntactic requirements and constraints of a construction which has been started, and being able to cope with these as the full plan unfolds. Assessment of the performance will involve, amongst other things, charting the relationships between fluency, accuracy, and complexity within the micro-planning units.

Recently, an area of research has opened up into the way speakers draw on memorized sequences of language (see for example Ellis 1996; Weinert 1995) rather than constructing streams of speech word by word. Such lexicalized sequences range from the invariable or formulaic (*you can't teach an old dog new tricks*; *be that as it may*) to the variable (NP be *two* NP *short of a* NP; NP be *the sort of person who would go around* V-ing NP) but all share the important characteristic that their surface complexity seems to owe less to syntactic processing than it does to simple selection from a memory store. Though this suggests that proficiency is not necessarily reflected in surface complexity of language, it is still valid to suppose that more proficient speakers are those who are able to keep track of where they are, syntactically, as they incorporate fully or partially fixed sequences with language freshly minted for the occasion.[2] Foster (in press) has shown that adult native speakers of English are adept at this and seem to rely more on memorized sequences of language when under some degree of communicative pressure than they do when the pressure is off. Non-native speakers on the other hand, seem to rely far more on word-by-word processing, increasing significantly in their complexity, fluency, and accuracy when not under communicative pressure (see also Foster and Skehan 1996). It is perhaps useful, therefore, in determining what characterizes the profile of a proficient speaker, whether of L1 or L2, to consider both the extent to which a person is able to exploit a store of native-like memorized sequences, and the extent to which a person is able to construct complex and extended syntactic turns out of smaller segments. In this paper we are concerned with measuring the second of these, acknowledging the importance of the first but leaving that problem of how to measure and investigate memorized language to others (see Foster in press; Willis 2000).

In deciding upon the most appropriate unit for segmenting oral speech samples, we must be governed by the well-established methodological criteria of reliability and validity. If the unit cannot be reliably identified, the measurements made will be misleading. Crookes (1990), in the first paper to address this issue in detail, noted that if the unit has little or no relationship to the psycholinguistic planning process, the measurements made will have little or no value. They will have no advantage over, say, merely chopping up the transcript into groups of ten words or twenty morphemes, or whatever. To put it simply, what we need to know is what the performer can achieve in a single chunk of micro-planning activity, and how particular types of plan may affect the complexity, accuracy, and fluency of the language that is produced.

It follows that if speech researchers can agree on a unit which is psycholinguistically valid and reliable in its application to speech samples, then it should, in theory, be possible to establish an international standard which would enable comparisons to be made across data sets, and ideally, across different languages. In practice, however, this would probably be difficult to achieve. We cannot enter into a prolonged discussion of the issues here, including that of the limitations on the applicability across languages of

the unit we identify in this paper, nonetheless its use should allow more reliable comparisons with research in other languages whose syntax is similar to that of English.

Two problems characterize the units of segmentation already in use by researchers working with spoken discourse and bedevil any attempt to make fruitful comparisons or replications. These are:

1. Definitions: ostensibly identical units are either defined in different ways, or not defined at all, or defined in a way which is too simple to be used with real spoken data.
2. Applications: if exemplified at all, definitions are accompanied by one or two citation examples which bear little resemblance to the messy reality of speech transcripts.

In the sections that follow we examine both of these problems in more detail.

2. DEFINITIONS: SIMILAR OR DIFFERENT?

Recently we looked at a significant and substantial sample of research published over the last seventeen years[3] in which either written or spoken language had been segmented in some way for quantitative analysis. Four leading journals in the applied linguistic and second language field were surveyed: *Applied Linguistics*, *Language Learning*, *Studies in Second Language Acquisition*, and *TESOL Quarterly*. Our survey revealed a plethora of definitions of units of analysis, a paucity of examples, and a worrying tendency for researchers to avoid the issue altogether by drawing a veil of silence over their methods. A total of 87 studies were identified which used some kind of unit to segment language for analysis. Of these 87, 44 provided no definition at all of the unit they chose to employ. The worrying implication in this omission is that there is no need for a linguist to define an 'utterance' or a 'T-unit', or any other unit of analysis chosen as a research tool: it is enough merely to label it. However, anyone involved in the analysis of spoken data knows that this is not in fact the case. In order to achieve both accurate and comparable analyses based on a quantitative evaluation of the data, you need more than a simple label to work from.

The following extract serves to illustrate how difficult it is to work with the fragmentary and elliptical data which is typical of oral language samples, particularly from second language learners. This is a typical piece of non-native oral interaction, taken from a research study in which participants had to discuss prison sentences for a list of offenders (Foster and Skehan 1996):

A: which which what is your opinion?
B: (1.0) maybe er (5.0) he (7.0)
A: long time? or it's for for you it's a major mistake or a small mistake?
B: maybe three months
A: three months for this one okay for me it's ten

B: ten?
A: ten years
B: yeah ten years oh very long

Without providing at least a definition of the unit employed to segment such data, is it possible for the reader to know how it should be dealt with? Is B's second turn counted as a unit even though it lacks a verb? Is there a unit in B's third turn, (which appears to be a clarification request) even though it is only one word? How many units are there in A's third turn? Or B's final turn? Can B's first turn be included in any way? In sum, how much of this is usable, and in what terms?

We found that 43 of the 87 studies in our survey did at least provide a definition of the unit they used. However, these varied considerably in the amount of detail provided, and studies which used the same unit often did not work from the same definition of it. In examining the kinds of definitions given, we would argue that they can be grouped into three broad categories: semantic, intonational, and syntactic.[4] We shall look at each of these in turn, giving examples of the definitions provided by the researchers who used them and discussing briefly the issues they raise.

2.1 Mainly semantic units

- *Proposition*: 'a semantic unit consisting of at least one major argument and one or more predications about this argument' (Sato 1988: 375).
- *C-unit* (semantic focus): 'utterances, for example, words, phrases and sentences, grammatical and ungrammatical, which provide referential or pragmatic meaning' (Pica *et al*. 1989: 72).
- *Idea unit* (semantic focus): 'a chunk of information which is viewed by the speaker/writer cohesively as it is given a surface form . . . related . . . to psychological reality for the encoder' (Kroll 1977: 85).

The identification of semantic units based on information/meaning chunks might initially appear appealing. However, it is clear that the extent of an idea, or an argument, is never easy and often impossible to establish with certainty. Thus definitions which rely exclusively on semantic criteria will tend to be extremely hard for the analyst to work with reliably. It is therefore understandable that in the literature these semantic criteria seldom stand alone: they are usually supported by grammatical and intonational ones.

2.2 Mainly intonational units

- *Tone unit/phonemic clause*: 'a distinctive configuration of pitches, with a clear centre, or nucleus. . . . The nucleus is the syllable (or, in some cases, series of syllables) which carries the greatest prominence within the tone-unit' (Crystal and Davy 1975: 16).
- *Idea unit* (intonation focus): 'The 'tone-unit' of Crystal . . . is essentially the same. . . . Most idea units end with an intonation contour that might

appropriately be called clause-final: usually either a rise in pitch, . . . or a fall. . . . A second factor is pausing. Idea units are typically separated by at least a brief pause' (Chafe 1980: 13–14).

- *Utterance*: 'a stream of speech with at least one of the following characteristics: 1. under one intonational contour; 2. bounded by pauses; 3. constituting a single semantic unit' (Crookes and Rulon 1985, cited in Crookes 1990: 187).

The central focus of each of these definitions is intonational. Chafe and Crookes both incorporate pausing as a secondary feature, and both also incorporate semantic criteria. However, for the second language researcher, units which rely purely or mainly on pausing and intonational features are particularly problematic because of the vagaries of such features in the speech of non-native-speakers. Pauses in L2 performance are not necessarily at unit boundaries, and it can be difficult to distinguish between pauses that result from message formulation or a lexical search. Also, the tone unit is likely to prove especially unhelpful in that dysfluent speakers are likely to produce more nuclei than fluent ones. Even fluent native speakers tend to produce a fairly large number of sub-clausal tone units. Such units are unlikely to reveal much about planning abilities and proficiency. Proponents of the third intonational unit, the utterance, such as Crookes (1990), do not seem always to distinguish it clearly from the idea unit, as the definition makes clear; in terms of intonational patterns, will the 'intonation contour' be a short mononuclear one, similar to that of the tone unit, or will it cover a longer stretch of speech, and be identified by a gradual declination of fundamental frequency? Intonational criteria are then, in our view, rather unstable.

However, although intonation may be an insecure single foundation for the unit definition, it may in addition, as Loban (1976) pointed out, be a useful complement to other criteria, especially syntactic ones. It is to syntactic units that we now turn, as we consider that these promise greater reliability and validity. Below we examine in rather more detail the range of syntactically based units which have been used in recent research. Before we do, however, it will be useful to define the subordinate unit common to each.

- *Clause and s-node*: 'either a simple independent finite clause, or a dependent finite or non-finite clause' (Foster and Skehan 1996: 310); 'S-nodes are indicated by tensed or untensed verbs' (Ellis *et al.* 1994: 483).

The terms 'clause' and 's-node' are interchangeable. They provide a relatively unproblematic measure to subdivide units into smaller segments. As we shall see, the problems arise when defining how speakers organize these smaller segments in longer and more complex syntactic units.

2.3 Mainly syntactic units

- *Sentence*. The sentence, as is widely recognized (see e.g. Quirk *et al*. 1985: 47) is problematic for spoken (and written) data, and will be ignored in this paper.
- *Idea unit* (structurally defined): Kroll gives a lengthy definition, which can be summarized as: An Idea Unit is a clause with its pre- and post-V clause elements. Also counted as IU's are non-finite subordinate clauses, and finite relative clauses where the relative pronoun is present (Kroll 1977: 90).
- *T-unit*: our survey revealed the T-unit to be clearly the most popular unit for the analysis of both written and spoken data. Hunt (1965, 1966, 1970) defines the T-unit as essentially a main clause plus any other clauses which are dependent upon it. Researchers using the T-unit and wishing to define it use one of Hunt's four versions.
  'one main clause with all subordinate clauses attached to it' (Hunt 1965: 20);
  'one main clause plus whatever subordinate clauses happen to be attached to or embedded within it' (Hunt 1966: 735);
  'the shortest units into which a piece of discourse can be cut without leaving any sentence fragments as residue' (Hunt 1970: 189);
  'a main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it' (Hunt 1970: 4).

These four versions suggest that 'non-clausal structures' and 'sentence fragments' may be included or excluded from an analysis depending upon which T-unit definition the researcher chooses to adopt.

As can be seen below, other definitions repeat Hunt's (1965, 1966, 1970) basic criteria, although Schneider and Connor (1990) allow a 'non-independent clause' to stand as a T-unit in written data if it is punctuated as a sentence.

> Schneider and Connor (1990: 427): 'a T-unit is 1. Any independent clause plus all its required modifiers, 2. Any non-independent clause punctuated as a sentence, 3. Any imperative' (for written data).
>
> Young and Milanovic (1992: 409): 'an independent clause and any associated dependent clauses' (for oral data).
>
> Santos (1988:73): 'a single independent clause together with all of its modifying subordinate clauses' (for written data).

There are indications in the literature that the T-unit definition is inadequate to deal with a full analysis of spoken discourse.[5] Tarone (1985) remarks that she was unable to analyse some of her recorded speech because it was so dysfluent, there were so few complete sentences (sic) and so much hesitation and repetition. Young, also working with oral data, and using the T-unit for analysis, defines this unit in the text unremarkably as 'one clause plus any subordinate clauses' (1995: 19). However, he adds a substantial endnote to

this simple definition, presumably because he found it insufficient for the task of segmenting his oral data:

> The following elements were counted as one t-unit each: a single clause, a matrix plus subordinate clause, two or more phrases in apposition, and fragments of clauses produced by ellipsis. Co-ordinate clauses were counted as two t-units. Elements not counted as t-units include back channel cues such as mhm and yeah, and discourse boundary markers such as okay, thanks or good. False starts were integrated into the following t-unit (Young 1995: 38).

These modifications move some way beyond the original idea of a unit as a main clause plus any subordinate clauses and one might wonder why these significant modifications are footnoted rather than included in the working definition in the text. With this footnote, Young is describing a unit very like the unit which came to be called a communication unit, or c-unit. This measure, first described by Loban in 1966, is an attempt to make the T-unit work with the elliptical nature of the spoken language. Loban first defined it in these terms:

- *C-unit*: 'grammatical independent predication(s) or . . . answers to questions which lack only the repetition of the question elements to satisfy the criterion of independent predication. . . . 'Yes' can be admitted as a whole unit of communication when it is an answer to a question such as 'Have you ever been sick?' (Loban 1966: 5–6).

Many, but by no means all, researchers investigating spoken language data have used the c-unit, though not necessarily Loban's definition of it. Pica *et al.* (1989: 72) extended Loban's definition into the definition we have classified as semantically-based above—a rather vague catch-all term which is not particularly illuminating or helpful. Lin and Hedgecock (1996), also using the c-unit for oral data, adopted Chaudron's (1988: 45) definition of a c-unit: 'an independent grammatical predication, the same as a T-unit except that in oral language elliptical answers to questions also constitute predication'. A major problem with these Loban-based definitions is their seeming exclusion of elliptical constructions which arise *within* a speaker's turn rather than link to an interlocutor's question.[6]

Brock (1986: 52–3) refers to Loban (1966) but surprisingly, she also adopts Freed's (1978) definition that a c-unit 'may have several sentence nodes as a consequence of having several sentences, several clauses, or being a run-on or compound sentence' (1978: 43). This definition, unlike any other for the c-unit, appears to describe a very long unit capable of encompassing several 'sentences' and is thus outside clear syntactic criteria.

The main choice amongst these units lies between those which are essentially clause-based (the clause itself, the S-node and the Idea Unit) and those which are potentially supra-clausal (the c-unit and the T-unit). While the clause-based units enable easier analysis in many—though not all—cases,

the potentially supra-clausal units offer greater validity. They allow the analyst to give credit to performers who can embed clauses and hence construct chunks of speech which reflect more sophisticated planning processes. However, coping with the different degrees and types of embedding and attachment in these longer units raises a range of problems. It is to these which we now turn.

3. APPLICATIONS: SILENCE AND SIMPLIFICATION

Occasionally, but not often, researchers exemplify the units they have employed with models of what they look like. For the most popular, the T-unit, the following examples have been provided.

> When the boy woke up he noticed the frog was gone (Cohen 1989: 148).
>
> He goes to the bookmaker and gets some money (one T-unit) (Lennon 1990: 406).
>
> When you make a milk shake, you mix it in a blender (one T-unit and 2 s-nodes) (Ellis *et al*. 1994: 483).

Schneider and Connor (1990) are exceptional in providing a whole section of their data with T-unit boundaries included, a part of which is given below:

> There are many different contributions between artists and scientists to society. || First artists contribute to society for entertainment. || Many people need it for relax after hard work. || Artists contribute to society as film artists, singers and so on. || Furthermore artists contribute to society with make new-work fields which are related with kind of activity. || (Schneider and Connor 1990: 415)

This is straightforward enough, and certainly the common basis for all these definitions (that a T-unit is a main clause with all its dependent/subordinate/modifying/associated clauses) works well here.

However, the reason it works well is, of course, that the data is written language. Transcribed spoken data is not nearly so tidy and clear cut. Application of T-units to the segmentation of oral data requires considerable modification of the unit. But our survey of recent research clearly showed many studies of speech have happily used T-units without admitting to any problems, raising the interesting question of what proportion of the data ended up, so to speak, on the cutting-room floor. Remaining silent on the subject, or only providing simple and unproblematic examples of how the T-unit was employed, shows researchers dodging some difficult but important questions.

To date, we have been unable to unearth any published examples of segmentation into c-units of an extended oral text. The emergence of a segment designed for oral data certainly seems to acknowledge the shortcomings of the classic T-unit, and is, therefore, to be applauded. However, the

lack of examples to illustrate what a c-unit might look like, especially one 'with several sentences', means we cannot be altogether certain what a c-unit analysis might involve. To address this problem, we now examine some typical features of oral performance to illustrate how difficult it is in practice to apply these rather spare c-unit definitions to the segmentation of oral data.

3.1 'Because' adverbial clauses

Optional adverbial clauses introduced by 'because', are frequently problematic in oral language. This is due to the fact that the relationship between the 'because' clause and its putative main clause is often uncertain, and the word appears to function as much cataphorically as anaphorically. Additionally, 'because' often performs a discourse marker function. In examples 1, 2, and 3 below 'because' could be paraphrased as 'I say this because . . .'. This discourse function is frequently signalled by pause and intonation phenomena.[7]

1 A: it's usual in this age to get in love with an older ` person because I'm talking about what happened to me (1.0) because you see experience in that older person.

2. A: I think I would ask er some- no er a ` judge.
   B: Yes!
   A: because she is quite in dangerous position.

3. B: . . . I would like to to see now a Europe a strong Europe like er maybe in the eighteenth century (1.0) because er I think we're going to have a lot of more more troubles erm in the next er decade (1.0) because er Japan for example or Russia and America are the two countries two strong countries and Europe is between the two these two countries.

The question here is whether the apparently dependent 'because' clauses should be accorded independent status.

3.2 Co-ordination

T-unit analysis, while it decrees that coordinated main clauses should be treated as separate units, also rules that coordinated verb phrases with the same subject should be treated as part of a single unit, as shown in the Lennon (1990: 406) example quoted above. In the following data however it is clear from pause and intonation phenomena, that the subsequent verb phrase constitutes a new start:

4. and the woman, um, (0.5) speaks um, um, go telephone, speaks telephone, (1.0) and look in in the other direction,

5. sometimes maybe bring some powder milk (1.5) and put it to some some er water and some nutrients protein or something like that this may be also sort of artificial milking.

6. The other woman is very happy now (0.5) and (3.0) just walking away with a gr great smile.

In many instances, non-native speakers will drop subjects in this way, creating apparent coordinated phrases which are almost certainly not intended as such.

3.3 'Topical' noun phrases

As Bygate (1988) has noted, independent noun phrase satellite units are common in speech, and particularly so in the case of second language learners whose first language is typologically a topic-comment language. Neither the T-unit, nor the c-unit, as previously defined, provides a satisfactory answer to the analysis of such phenomena. The problem is exemplified in the underlined sections of 7, 8, and 9 below.

7. It's just a matter of passing (exams) And especially the basic education (0.5) they have to pass automatically from one grade to another

8. Apples are grown in Pakistan These fruits are same as like yours but not gooseberries and strawberries these are not grown in my country

9. and on the pool three people maybe they are children three people they are swimming now

3.4 Scaffolding and interruption

Co-operative building up of a conversation, with sharing of units, is common in highly interactive conversations. However, the analysis of such interactions raises quite complex questions about how the resulting units are to be divided up and assigned. Examples 10, 11, and 12 are typical:

10. B: I think um (3.0) yeah it's hard maybe he like her but no like a girlfriend, just a
    A: Just like a sister.
    B: Yeah.

11. A: They both look fairly sort of er
    B: miserable
    A: yeah
    B: for a wedding day

12. A: This guy looks like he is some kind of a (1.0)
    B: priest?
    A: priest. He might be an ayatollah

Given the breadth and range of articles in which the detailed analysis of oral data has been required (as noted above, we found many in our survey of just four journals), it is clear that other researchers must have faced similar

problems to those outlined above. The dearth of published information about how their analyses were achieved is frustrating. Oral data is particularly messy; second language oral data is generally messier than first language data. Researchers who face the difficult task of segmenting it for analysis are presently doing so largely without the benefit of the shared experience of their fellows. What is needed is an accessible standard unit of analysis, explicit and exemplified, which is psycholinguistically valid, and which can be applied reliably to a wide range of oral data. In the next section therefore we present a detailed definition of such a unit, illustrated with examples from authentic oral data, with justification for our choices.

4. THE ANALYSIS OF SPEECH UNIT (AS-UNIT)

The unit we have adopted, and termed the AS–unit, is a mainly syntactic unit. There are a number of reasons we consider this to be a valid unit for analysing spoken language. First, the studies of pausing in native-speaker speech suggest that syntactic units are genuine units of planning, since many pauses occur at syntactic unit boundaries, and especially clause boundaries (e.g. Raupach 1980; Garman 1990). Secondly, our definition allows analysis of speech units which are greater than a single clause since there is evidence from intonation and pause features that speakers may plan multi-clause units (Beattie 1980). We propose that the ability to plan at the multi-clause level is important for establishing a speaker's level of proficiency, and evaluating the complexity of a particular performance by that speaker. There is a good prima facie case for interest in such a unit since the ability to produce units with more than one clause seems to be associated with planned speech (e.g. Foster and Skehan 1996), and with L2 development in instructed students (Tonkyn 1996).

Our unit takes Hunt's T-unit as its starting point and then elaborates this to deal with the features characteristic of spoken data.

4.1 Definition of the unit

An AS-unit is a single speaker's utterance[8] consisting of *an independent clause, or sub-clausal unit*, together with any *subordinate clause*(s) associated with either. In the examples that follow below an AS-unit boundary is marked by an upright slash . . . | . . . A clause boundary within an AS-unit is marked by a double colon (::). False starts, functionless repetitions, and self-corrections are put inside brackets {..}.

An independent clause will be minimally a clause including a finite verb.

13. | That's right |

14. | Turn left |

15. | I take a different way |

16. | You go to the main street of Twickenham |

However, unlike the T-unit, our definition allows for the inclusion of independent sub-clausal units, which are common in speech, and specifies the nature of these more clearly than has previously been the case where the c-unit has been used as the unit of analysis (see (b) below). Our unit is still primarily syntactic because this offers an acceptable level of reliability, given that syntactic units are easier to identify than intonational and semantic ones. However, in addition to this we argue for the principled use of intonation and pause phenomena to deal with some awkward cases, and we discuss these below, together with examples from native speaker and non-native speaker data which we have been collecting and analysing.

An independent sub-clausal unit will consist of: *either* one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation:

17. A: | how long you stay here |
    B: | three months. |

or a minor utterance, which will be defined as one of the class of 'Irregular sentences' or 'Nonsentences' identified by Quirk *et al*. (1985: 838–53.)

18. | Oh poor woman |

19. | Thank you very much |

20. | Yes |

A subordinate clause will consist minimally of a finite or non-finite Verb element plus at least one other clause element (Subject, Object, Complement or Adverbial).

21. | I serves in in a organization government organization in Bangladesh :: which is called er department of agricultural extension | (2 clauses, 1 AS-unit)

22. | I have no opportunity to visit | (1 clause, 1 AS-unit)

23. | it is my hope :: to study crop protection | (2 clauses, 1 AS-unit)

24. | and you you be surprise :: how he can work | (2 clauses, 1 AS-unit)

Since subordination is frequently used as a measure of complexity (see, e.g., Crookes 1989; Foster and Skehan 1996; Wigglesworth 1997) it is crucial to clarify the issue of what constitutes a subordinate clause. In particular, the classification of non-finite clauses can be difficult. For example, in the utterance: 'I like reading', 'reading' could be analysed as a minimal non-finite clause, or simply as an NP. In this analysis, we require at least one additional clause element to establish clausal status, as in (23) above.

Clause coordination and subordination require additional clear specification. First, there is the problem of coordinated Verb Phrases, as exemplified in examples 4, 5, and 6 above. We argued that in many cases, the second VP,

even if the subject is omitted, constitutes a fresh start for the speaker as shown in example 6. To address this problem we have identified conditions for determining where the unit boundary lies.

In cases where coordination of verb phrases occurs, the coordinated phrases will normally be considered to belong to the same AS-unit, unless the first phrase is marked by falling or rising intonation and is followed by a pause of at least 0.5 seconds. This represents a clearly noticable pause which can be reliably measured without access to specialist equipment.

25. | and they pinned er a notice to his front :: telling everybody :: what he had done (0.5) | and marched him around the streets with a gun at his back | (2 AS-units)

Subordinate clauses within an AS-unit can realize the following functions:

(a) subject (initial or postponed)

26. | sometimes it creates problems :: that he knows nothing |

(b) verb complementation (object, complement, or catenative verb complementation)

27. | and er they told :: that there there was no food crisis |

28. | I wish :: to er visited other areas of England |

29. | I would like :: to ask you :: if you can give me three weeks leave now |

30. | the main object of this organisation is :: to raise up the people's attitudes |

(c) phrasal post-modifier or complement

31. | still in our country the school and er college students learned the English :: which were er taught to the students before thirty years. |

Under certain conditions, the subordinate clause within an AS-unit can realize an adverbial function. Initial and medial examples do not usually cause a problem: it is clear where they belong as shown in examples 32 and 33:

32. | when I was in the university :: er I have specialized in this er subject |

33. | very early in my university when I was study :: I I did my thesis my underground thesis in pasture grassland |

However, the loose concatenations of optional subordinate adverbial clauses, particularly in final position, do cause problems as discussed in examples 1 to 3 above. We have, therefore, established a fairly tight condition for allowing such clauses to be included in the preceding AS-unit.

Specifically, such a final adverbial clause should be within the same tone unit as at least one of the other preceding clause elements of the AS-unit.

We see this linkage as establishing that the adverbial clause is definitely part of the plan which produced the initial main clause.

34. | and I can bring him tomorrow together :: where you can talk with him | (1 AS-unit)

35. | I can under'stand :: when I 'read scien ` tific English | (1 AS-unit)[9]

36. | specially for 'reading scien'tific ` papers | because er all the 'papers that er ar'rived to the 'library in 'Chile are ` English paper | (2 AS-units)

4.2 False starts, repetitions, and self-corrections

Under certain circumstances, it may also be necessary to consider how certain dysfluency features, such as false starts, repetitions, and corrections, will be handled in relation to the unit. These are frequent phenomena in oral language data, and particularly so in second language data. Different researchers may wish to deal with actual linguistic material within the false starts and corrections in different ways, depending on their interests. For example, if words per unit are being calculated as a measure of complexity it is necessary to have a principled way of excluding such phenomena from the total word count. We define a false start as an utterance which is begun and then either abandoned altogether or reformulated in some way:

37. | {That's cos } you're saying that :: cos you're a man |

38. | {What about } can you give me a credit slip? |

In cases where an AS-unit is produced before the message is abandoned, that part of the utterance which meets the AS-unit criteria will be counted as an AS-unit, with the remainder being recorded as a false start.

39. | I have to do this project { because in Brazil } |
| is completely different here |

A repetition is where the speaker repeats previously produced speech. This is a device which may be used to hold the floor, or to allow time for planning on line. However, it is necessary to distinguish between those repetitions which indicate dysfluency, and those which are used for rhetorical effect.

40. | { and one person } er one person enjoy the canoe |

41. | it's a very very bad man |

A self-correction occurs when the speaker identifies an error either during or immediately following production and stops and reformulates the speech; self-corrections will therefore include an element of structural change:

42. | I think ::{ they're a very } they have good time |

43. | I'll work out :: {what should I } what can we do for you |

The following example includes both a self-correction, and a repetition:

44. | so he { decides }{ decided } decided :: to go fishing |

Where a self-correction occurs, the final version is counted as an AS-unit, with previous versions excluded.[10]

4.3 Topicalization

Topicalized noun phrases generally belong to the unit of which they are the topic; thus the following comment includes the (generally pronominalized) noun phrase repetition of the topic.

45. | but modern office { they has } they had er many convenient machine |

46. | and some children they are playing the ball |

However, noun phrase satellite units which are separated from the following AS-unit by falling intonation *and* a pause (equal to or greater than 0.5 sec) will be treated as separate AS-units. Thus the underlined section of example 7 above would be treated as a separate AS-unit because the word 'education' is marked by falling intonation and followed by a pause of 0.5 seconds.

4.4 Interruption and scaffolding

In highly interactive discourse, interruption and scaffolding are common and can present considerable problems for analysis as indicated above. A number of these problems can be addressed by adopting one of the different levels of analysis outlined in the next section. Others, however, require specific discussion. In example 47, the subordinate clause is latched to the preceding utterance by the same speaker, but speaker B has interrupted. In this case, A's utterances would be analysed as one AS-unit with 3 clauses provided that the previously stated criteria for including a final adverbial in the preceding AS-unit are met.

47. A: | oh that's a big problem
    B: oh no!
    A: :: because my shop's policy is only :: to give the credits for the return goods |

In the following example, the first speaker (who is a seven-year old child relating a story from a picture book) cannot access the collective term for bees, which is provided by her mother (B). In this case, speaker A is credited with two AS-units in the following example, with the second including a repair which is concluded after the interjection by speaker B:

48 A: | the dog pushed the tree |
   | { the the herd of bees fell } the um
   B: | hive the hive of bees|
   A: hive of bees fell down |

In example 10, above, 'just like a sister' completes the AS-unit of B, and is also credited to speaker A, as a complete AS-unit.

Finally, we return to the extract we began with, now coded for AS-units:

A: | { which which }what is your opinion? |
B: (1.0) { maybe er } (5.0) { he } (7.0)
A: | long time? |
| { or it's for } for you it's a major mistake or a small mistake? |
B: | maybe three months |
A: | three months for this one |
| okay |
| for me it's ten |
B: | ten? |
A: | ten years |
B: | yeah |
| ten years |
| oh very long |

4.5 Three levels of application

Up to now, researchers have often felt the need to exclude certain data from their analyses, but this has frequently been done on what appears to be an ad hoc basis, with no explanation. The following is an attempt to provide criteria for the principled exclusion of data where this is necessary for the purposes of coherent analysis. Thus the three different levels of inclusiveness are proposed to cope with a range of research purposes and different types of spoken language data. To clarify further how these levels are intended to work we have included an appendix with an example of a transcript analysed under the three different levels.

Level One: to be used for a full analysis of all the data.

- Include everything except untranscribable data, although single inaudible words of identifiable word class should be included.

Level Two: to be used for highly interactional data. This is for researchers who are working with interactional data which can yield a high proportion of minimal units (e.g. one-word minor utterances and echoic responses) whose inclusion in an analysis could distort the perception of the nature of the performance.

- Exclude one-word minor utterances:

48. Yes; No; Okay; Uhuh; Right

- Exclude echo responses which are verbatim:

49. A: I think two years
B: Two years

Level Three: This is for use in special cases where analysis of non-fragmentary AS-units is required. For example, in cases where performances on differing types or sections of OPIs[11] need to be standardized. This level

takes the principle behind level 2 a stage further. It is principally for researchers who are primarily interested in what the performer can do in the production of relatively 'complete' units. In this case, largely transactional data may be required, but interactional interludes may produce, for some subjects, uncharacteristically short units whose inclusion in the analysis could distort any comparison of subjects. This level introduces a greater degree of standardization of the unit, without eliminating elliptical units altogether.

- Exclude items in level 2 above. In addition:
- Exclude V-less elliptical AS-units involving ellipsis of elements of the interlocutor's speech.

50. I: what is your mother tongue then
    A: Arabic Arabic

- Exclude AS-units involving substitution of clause, predicate, or predication level units of interlocutor's speech.

51. I: yes but do you find that there are periods say over a period of ten years when the weather is drier and then over the next ten years it may be wetter do you find that as well
    L: yes I think so

- One or two word greetings and closures can also be excluded in this level of analysis.

5. CONCLUSION

Clearly, neither the definition, exemplification, or the employment of units of segmentation for oral data is a straightforward matter. Anyone embarking on an analysis of oral data must quickly realize that numerous decisions have to be made as to how to deal with problematic data which will not easily fit the simple definitions of units currently available. While researchers spend considerable time, toiling in the dark, making decisions about how to segment their data using the unit they have chosen, such detail cannot usually be included in the research report due to the restriction on space to which published articles are subject, and thus other researchers may not be able to benefit from their insights.

The purpose of this paper is therefore twofold. In the first place we have sought to identify and discuss the shortcomings of the units of segmentation currently in widespread use in oral data analysis. We have tried to show that even where definitions are provided and examples are given these are inadequate to deal with the fragmentary nature of oral data.

In the second place we have sought to provide a useful resource for the researcher. We have presented a comprehensive definition of a unit which is accessible, clearly defined and easily applied—criteria which the AS-unit

meets. In addition, it is capable of the flexibility required to suit a variety of research purposes, hence the 'unit for all reasons' of our title. We have argued that the unit is valid, and sensitive to genuine differences in performance. We have also provided extensive examples over a wide range of data to assist researchers in the segmentation of their data and we hope that this detail in our definition will mean that other researchers will be able to use the unit reliably. As a consequence perhaps, second language researchers with an interest in oral language data can start to talk the same language.

Revised version received January 2000

ACKNOWLEDGEMENTS

We are grateful to three anonymous *Applied Linguistics* reviewers for their very helpful and insightful comments.

APPENDIX

The following example transcripts are analysed into AS-units, with each unit beginning on a new line. All units would be included in a first level analysis. Units which would be excluded from a second level analysis are indicated in bold plain type. Additional units which would be excluded from a third level analysis are indicated in bold italic type. False starts are inside brackets { . . . } To summarize:

Level 1 = all units
Level 2 = excludes units in **bold plain type**
Level 3 = excludes units in **bold plain type** *and bold italic type*

A: | **oh yes** |
| **hello** |
B: | **hello** |
| can I help you? |
A: | ***yes I hope so*** |
| I bought this cassette recorder here last weekend |
B: | **yes** |
A: | and when I took it home :: I couldn't record with it |
B: | **oh** |
A: | so er I'm really upset |
| and I want my money back please. |
B: | **oh yeah** |
| ***oh very sorry*** |
| { I think } I apologise for the faulty goods |
A: | **oh right** |
| **well** |
| **yes** |
B: | this shops { is prol is er only give } can't give er the money back |
A: | you can't give my money back? |
B: | **yeah** |
| ***very sorry*** |

A: | but it's no good |
| I don't want another one really |
B: | oh { this er } have you got a receipt? |
A: | { well actually } I'm sorry |
| no I can't find the receipt |
| I've looked everywhere |
B: | Ah ah I don't know er |
| I thinks :: er need to go :: to er ask the manager |
A: | oh well yes perhaps I should yes |
B: | yes because you haven't got a receipt |
A: | **mm yes** |
B: | { I ca } I can't you give money back |
A: | you can't |
A: | oh I see |
A: | **oh right** |
B: | I'm really sorry |
A: | { what about } can you give me a credit slip? |
B: | **er credit slip?** |
| er you haven't got { a s } a receipt |
A: | **receipt no** |
B: | er ask the manager first |
| I think :: it's better |
A: | **oh okay** |
| **alright** |
| ***I suppose so*** |
| you're only doing your job |
B: | **yeah** |
A: | **alright** |
| when can I see the manager please? |
B: | **yes so** |
A: | now? |
B: | **yeah** |
A: | **oh good** |
| **alright** |
B: | **thanks** |
A: | **good** |
| **okay** |

NOTES

1 LARSP: Language Assessment, Remediation and Screening Procedure.

2 Memorized sequences of words cannot be merely threaded together, like beads on a string, in order to produce an extended utterance. The syntax of a sequence is often governed by what precedes it, or follows it. Consider the following, in which the first sequence constrains the verb form of the other three: '*This is not the time for turning a blind eye, or sticking your head in the sand, but for standing up to be counted*'.

3 A period of time corresponding to the resources available in the libraries we consulted.

4 It should be stressed that these three categories are broadly defined and features of the other categories may be taken into account when evaluating or defining units in a particular category. We acknowledge

that it is impossible completely to separate semantic, intonational, and syntactic criteria.

5 This is scarcely surprising as it was developed originally by Hunt to measure syntactic development in the written work of schoolchildren. What is surprising is that it should ever have been adopted for analysing interactive oral data.

6 We shall deal with one common form of these, the topical noun phrase, in some detail below.

7 In these and the following examples the number of seconds in a pause is given in brackets. ` indicates a falling intonation. Intonational diacritics are included only in examples where intonational criteria are being appealed to.

8 The common phenomenon of scaffolding is discussed below in example 47.

9 ↑ = pitch 'hitched up'; ' = level stress; ` = falling nuclear tone; again, intonational diacritics are included only in examples where intonational criteria are being appealed to.

10 If the length of AS-units is being counted, false starts, or items which are replaced for grammatical or lexical reasons, will typically not be included in the count. The same would be true of functionless repetitions. However, the instances of reformulation and repetition may be noted, to establish, say, the fluency characteristics of the speech.

11 OPI = Oral Proficiency Interview.

REFERENCES

Beattie, G. 1980. 'The role of language production processes in the organisation of behaviour in face-to-face interaction.' in B. Butterworth (ed.): *Language Production: Vol 1: Speech and Talk*, pp. 69–109. London: Academic Press.

Brock, C. 1986. 'The effects of referential questions on ESL classroom discourse.' *TESOL Quarterly* 20/1: 47–59.

Brown, R. 1973. *A First Language: The Early Stages*. London: Allen & Unwin.

Butterworth, B. 1980. 'Evidence from pauses in speech' in B. Butterworth (ed.): *Language Production: Vol 1: Speech and Talk*, pp. 155–77. London: Academic Press.

Bygate, M. 1988. 'Units of oral expression and language learning in small group interaction.' *Applied Linguistics* 9/1: 59–82.

Chafe, W.L. 1980. 'The deployment of consciousness in the production of narrative' in W.L. Chafe (ed.): *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, NJ: Ablex, pp. 9–50.

Chaudron, C. 1988. *Second Language Classrooms: Research on Teaching and Learning*. Cambridge: Cambridge University Press.

Cohen, A. 1989. 'Attrition in the productive lexicon of two Portuguese third language speakers.' *Studies in Second Language Acquisition* 11/2: 135–49.

Crookes, G. 1989. 'Planning and interlanguage variation.' *Studies in Second Language Acquisition* 11: 367–83.

Crookes, G. 1990. 'The utterance and other basic units for second language discourse analysis.' *Applied Linguistics* 11: 183–99.

Crookes, G. V. and **K. Rulon.** 1985. *Incorporation of Corrective Feedback in Native Speaker/Non-Native Speaker Conversation*. Technical Report No. 3. Center for Second Language Research, Social Science Research Institute, University of Hawaii: Honolulu.

Crystal, D. C. and **D. Davy.** 1975. *Advanced Conversational English*. London: Longman.

Crystal, D., P. Fletcher, and **M. Garman.** 1976. *The Grammatical Analysis of Language Disability*. London: Edward Arnold.

Ellis, N. 1996. 'Sequencing in SLA: phonological memory, chunking, and points of order.' *Studies in Second Language Acquisition* 18: 91–126.

Ellis, R., Y. Tanaka, and **A. Yamazaki.** 1994. 'Classroom interaction, comprehension, and the acquisition of L2 word meanings.' *Language Learning* 44: 449–91.

Foster, P. in press. 'Rules and routines: a consideration of their role on the task-based language production of native and non-native speakers' in M. Bygate, P. Skehan, and M. Swain (eds): *Language Tasks: Teaching, Learning and Testing*, London: Longman.

Foster, P. and **P. Skehan.** 1996. 'The influence

of planning on performance in task based learning.' *Studies in Second Language Acquisition* 18: 299–324.

Freed, B. F. 1978. Foreigner talk: A study of speech adjustments made by native speakers of English in conversation with non-native speakers. Unpublished doctoral dissertation, University of Pennsylvania.

Garman, M. 1990. *Psycholinguistics*. Cambridge: Cambridge University Press.

Hunt, K. 1965. *Grammatical Structures Written at Three Grade Levels*. Champaign, Ill: National Council of Teachers of English.

Hunt, K. 1966. 'Recent measures in syntactic development.' *Elementary English* 43: 732–9.

Hunt, K. 1970. *Syntactic Maturity in Schoolchildren and Adults.* Monograph of the Society for Research into Child Development.

Kroll, B. 1977. 'Combining ideas in written and spoken English: a look at subordination and co-ordination' in E. Ochs and T. Bennett (eds.): *Discourse across Time and Space*. SCOPIL (Southern California Occasional Papers in Linguistics) no. 5.

Lennon, P. 1990. 'Investigating fluency in EFL: a quantitative approach.' *Language Learning* 40/3: 387–417.

Lin, Y.-H. and **J. Hedgecock.** 1996. 'Negative feedback incorporation among high-proficiency and low-proficiency Chinese-speaking learners of Spanish.' *Language-Learning* 46/4: 567–611.

Loban, W. 1966. *The Language of Elementary School Children.* (Research Report no. 1). Champaign, Ill: National Council of Teachers of English.

Loban, W. 1976. *Language Development: Kindergarten through Grade Twelve.* Urbana, Ill: National Council of Teachers of English

Pica, T., L. Halliday, N. Lewis, and **L. Morgenthaler.** 1989. 'Comprehensible outputs as an outcome of linguistic demands on the learner.' *Studies in Second Language Acquisition* 11/1: 63–90.

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

Raupach, M. 1980 'Temporal variables in first and second language production' in H. W. Dechert and M. Raupach, (eds.): *Temporal Variables in Speech: Studies in Honor of Freida Goldman-Eissler*. The Hague: Mouton.

Richards, B. 1987. 'Type-token ratios: what do they really tell us?' *Journal of Child Language* 14: 201–9.

Santos, T. 1988. 'Professors' reactions to academic writing of NNS students.' *TESOL Quarterly* 22/1: 69–90.

Sato, C. 1988. 'Origins of complex syntax.' *Studies in Second Language Acquisition* 10/3: 371–95.

Scarborough, H. 1989. 'Index of productive syntax.' *Applied Psycholinguistics* 11: 1–22.

Schneider, M. and U. Connor. 1990. 'Analysing topical structure in ESL essays.' *Studies in Second Language Acquisition* 12/3: 411–27.

Snow, C. 1996. 'Change in child language and child linguists' in H. Coleman and L. Cameron (eds.): *Change and Language.* Clevedon, Avon: Multilingual Matters.

Tarone, E. 1985. 'Variability in interlanguage use: a study of style-shifting in morphology and syntax.' *Language Learning* 35/3: 373–405.

Tonkyn, A. 1996. 'The oral language development of instructed second language learners: the quest for a progress-sensitive proficiency measure' in H. Coleman and L. Cameron (eds.): *Change and Language*. Clevedon: Multilingual Matters, pp. 116–30.

Wagner, K.R. 1985. 'How much do children say in a day?' *Journal of Child Language* 12: 475–87.

Weinert, R. 1995. 'The role of formulaic language in second language acquisition: A review.' *Applied Linguistics* 16/2: 181–205.

Wigglesworth, G. 1997. 'An investigation of planning time and proficiency level on oral test discourse.' *Language Testing* 14/1: 101–22.

Willis, J. 2000. 'Lexical chunks: identifying frequent phrases, classifying and teaching them.' Paper presented at Association for Language Awareness 2000 conference, University of Leicester 26–29 June.

Young, R. 1995. 'Conversation style in language proficiency interviews.' *Language Learning* 45/1: 3–45.

Young, R. and **N. Milanovic** 1992. 'Discourse variation in oral proficiency interviews.' *Studies in Second Language Acquisition* 14/4: 403–24.