UNITS OF PRODUCTION IN WRITING: TOWARDS A DISCOURSE PERSPECTIVE

Kristyan Spelman Miller
University of Reading

ABSTRACT

The identification of units in the production of language has a long history in both applied linguistic and psycholinguistic research, in particular where there is a concern for the description and measurement of productivity and fluency. Such applications in both L1 and L2 contexts have tended to focus on spoken language production, however, and to use principally syntactic, semantic or phonological criteria as the basis for the identification of units. The purpose of this paper is to extend and develop the notion of units of production to the case of written language, and to explore the possibility of a discourse-oriented unit of analysis of productivity and fluency in writing. Drawing on a study of L1 and L2 on-line writing processes which uses the recording of keystroke presses on a word processor to elicit fine-grained temporal records of production, we present a framework for the description and analysis of on-line written text production, focussing on the 'framing device' as a potentially useful unit in the analysis of text production.

KEY WORDS: Units of production, on-line writing processes, framing device, written language, topic.

RESUMEN

La identificación de unidades en la producción del lenguaje tiene una larga historia en los trabajos de investigación tanto de lingüística aplicada como de sico-lingüística, en particular en aquellos casos en los que existe interés en la descripción y medida de la productividad y la fluidez. Sin embargo, tales aplicaciones en contextos tanto de la L1 como de la L2 han tendido a centrarse en la producción de la lengua hablada, y en el uso de criterios sintácticos, semánticos o fonológicos, principalmente, como base para la identificación de unidades. El objeto de este artículo es ampliar y desarrollar la noción de unidades de producción al caso de la lengua escrita, así como explorar la posibilidad de una unidad orientada hacia el discurso en el análisis de la productividad y fluidez por escrito. Basándonos en los datos de un estudio sobre los procesos de redacción en línea en la L1 y la L2 que utiliza el registro del número de pulsaciones en un procesador de texto para elicitar registros temporales de la producción de grano fino, presentamos los resultados relativos a la productividad y tasa de producción de dos tareas de redacción académicas, y discutimos la importancia potencial de una unidad de discurso a la que aludiremos como 'marcador discursivo' ['framing device'] en el proceso de producción del texto escrito.

PALABRAS CLAVE: Unidades de producción, procesos de redacción en línea, marcadores discursivos, lenguaje escrito, tópico.

1. INTRODUCTION

An essential step in the description of language output in both psycholinguistic and applied linguistic research paradigms is the definition and identification of a basic unit of production. It is on the basis of such a unit that language performance and development may be described and quantified, allowing observations to be made of quantity of language produced, frequency and ratio of features, and calculations of such dimensions as accuracy, complexity and fluency. Given the traditional focus of language production research on spoken language, we are familiar with a range of units applied to the description and measurement of oral data. In a recent article, Foster et al (2000) provide a useful overview of a number of contenders, such as utterance, c-unit, tone-unit, T-unit, idea unit, and so forth, each regularly employed in studies of production, but often without clear definition and justification. While 'a unit for the segmentation of oral data is an essential tool in applied linguistics' (Foster et al 2000: 354-55), too little attention appears to be drawn to the importance of this choice, and to consequences for the comparability of research findings across the discipline.

The study of written language production is less widely documented and discussed than spoken language within the literature, and from many points of view, writing production remains 'a stepchild of psycholinguistics' (Bonin & Fayol 1996: 145). The increased interest in recent decades, however, in the study of writing from the perspective of real-time process rather than of finished product, has generated a substantial body of research on cognitive aspects of writing within a range of first language (L1) and second language (L2) contexts, but still little has been discussed of the textual aspects of production. What characterises the language as it is produced in real time? What units are appropriate for its description and analysis? It is this cognitive-textual dimension of writing which inspires this current paper.

The approach which we present here is one designed for the linguistic/textual analysis of written language produced in real time. The study to which it is attached is more fully documented elsewhere (Spelman Miller 1999, 2000a, 2000b), but in brief, constitutes an investigation of the writing of subjects (both L1 and L2 writers) using a word processor. A relatively novel data elicitation method, called keystroke logging, is used to record the writing activity of these subjects. This is an unobtrusive recording tool which logs all keyboard activity (that is, all presses of character and function keys and cursor movements) made by the subject as they write, and stores the keystroke data electronically as logfiles. This allows access to fine-grained information concerning the timing and sequencing of actions, and permits the detailed investigation of temporal aspects of the writing process: when, where and for how long writers pause during production of the text in order, among other things, to plan and revise. At the same time, of course, the text produced between pauses is also readily available for analysis.

This on-line perspective of the writing process opens up questions for the analyst concerning the nature of the text produced, and the best means to characterise the strings, or units, of language produced. When text is produced in real

time, what spans of text emerge? Do they conform to (written) grammatical units such as word, phrase, clause, sentence, paragraph level constituents, or are other characterisations (for example, discourse related units) feasible and interesting?

The goal of the paper which follows is to present a new framework developed to describe on-line textual data. Our discussion begins with an illustration of the type of keystroke data we are dealing with, and then briefly outlines the range of pause-related phenomena which arise from these data. Central to our analysis of on-line production is the definition of the location of pauses. The on-line, dynamic nature of our data prompts us to develop a flexible scheme for the identification of pause locations, initially through reference to grammatical elements within the text, and then, by extension, by introducing an alternative, exploratory unit referred to as the *framing device*. This offers a means of interpreting the potential function of certain units in the production of text in establishing and maintaining topic in the discourse. With reference to data in our study, we define and illustrate a number of proposed categories of framing device, and consider their importance in the description of units of written text production.

2. THE COMPUTER-BASED OBSERVATION OF WRITING

The decision to study the cognitive processes of writing opens up a number of options in terms of data collection methods. Studies in recent decades have frequently favoured approaches such as think-aloud verbalisation and retrospective interview, which are direct observation methods in eliciting data directly from the writer him/herself. In the case of the former, verbalisation technique, which involves the writer in producing verbal commentary on his/her thoughts and actions as the task is being performed, data are elicited in real time (synchronously), whereas retrospective interview is asynchronous, or outside the time frame of the composing activity. Both methods have strengths and weaknesses which are well documented in the literature (see Cooper & Holzman 1983, Ericsson & Simon 1984, Greene & Higgins 1994, Kowal & O'Connell 1987, Russo et al 1989, Smagorinsky 1994, for example). For our purposes, the decision was made to opt for an alternative synchronous procedure to verbalisation. Pause analysis, involving the observation of writing activity in real time, avoids the problem of reactivity (the writer producing speech at the same time as writing), but is an indirect observation approach, offering only indirect evidence of potential cognitive activity underlying overt behaviour. The advantage, however, is that it offers a means of collecting a mass of detailed information providing insight into the dynamic relationship between planning and formulating processes.

The availability of an unobtrusive computer-based recording technique for observing the writing event allows for a high level of detail in the observed data, which would not be feasible through less sophisticated video-recording (used, for example, in the seminal studies of Matsuhashi 1981, 1982, 1987, and Gould 1978, 1980). As the writer composes on the word processor, resident software records all operations made during the writing event, and stores information electronically in

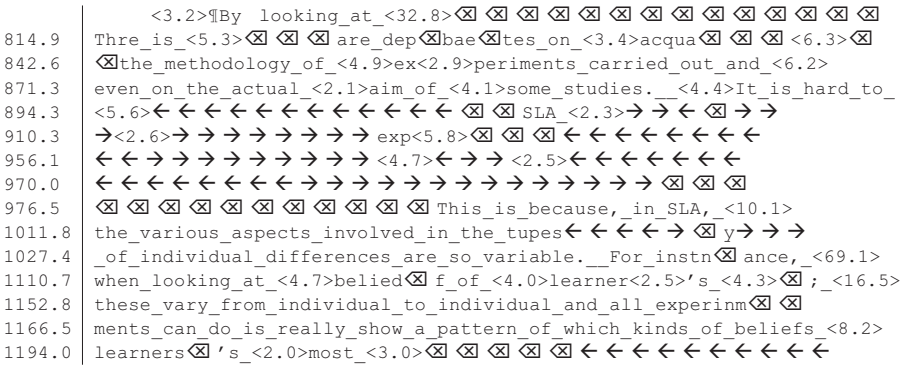

Figure 1. Sample from a keystroke log-file

logfiles for subsequent analysis. The output of the logfiles is a highly detailed record of the writing event, as illustrated in the text sample above (Figure 1). JEdit, the keystroke logging software used in our research (see Severinson Eklundh & Kollberg 1996 for full details) presents pausological information as bracketed numbers (e.g. <3.2> representing a pause of 3.2 seconds), cursor movements as left, right, up or down arrows, backspace deletion as the crossed arrow symbol, space bar presses as the underbar symbol (__), and all other characters (letters and punctuation marks) as produced in the text. The running time of the writing event (in seconds) is given down the left-hand margin.

3. PAUSE-RELATED PHENOMENA

The automatic logging of keypresses allows us access to a wealth of measurable data, from which observations may be made concerning a range of pause-related phenomena. Among these are the following:

– *pause duration*: instances of keyboard inactivity (greater than two seconds) measured in seconds and tenths of seconds;
– *pause frequency*: instances of inactivity identified between keyboard activity such as character presses or operations (such as deleting, scrolling);
– *pause rate*: the total number of words produced during the session divided by the total number of pauses;
– *productivity (length of text span)*: the length of text produced between two pauses (see discussion below);
– *rate of production (within a text span)*: the average number of words produced within the text span per minute, including time spent in producing keystrokes and pauses less than two seconds in duration.

For our discussion of units of production, the notion of the *text span* is of particular importance. Text spans between pauses may consist of units of different

lengths: single words, multi-word sequences, no text, or part-words (character sequences or single characters which do not constitute a word). From the logfile sample given in Figure 1, we illustrate the range of such units (listed below for ease of reference) generated in the production of the first sentence of the text:

By looking at < >[1]
There is < >
are debates on < >
acqua < >
the methodology of < >
ex < >
periments carried out and < >
even on the actual <>
aim of <>
some studies. <>

In the application of these text span categories to our data, a number of operational decisions were taken. Firstly, the category 'no text' refers to all spans in which no text is produced (for example, when scrolling takes place marked by cursor movement, when text is deleted, or when there is insertion of a punctuation mark but no text). Secondly, the character or 'part-word' category is distinguished from 'single word' through context. When a text span comprises two part words, as in example 1 below, this is categorised as a part-word span:

1 <>*cond lang*<> *(second language)*

Thirdly, cardinal numbers, for example in a list of points, are classed as a single word span, as are items such as *L2* and *SLA.* A pause within such an item is treated as word-internal, therefore. In some cases a text span may comprise a single word plus part-word. In this case, only the single word is recorded.

Productivity measures are calculated on the basis of the number of intact, recognisable words (that is, single words and multi-word sequences) produced between two pauses. Any whole words produced within the text span and subsequently deleted are included in the analysis. In example 2, therefore, the text span comprises three words (of which one, *is*, was at some point deleted):

2 <> *[is] has been* <>

Part-words deleted within the text span are not included in the productivity count. In example 3, for example, the text span consists of eleven words, since the deleted items (marked in square brackets) do not comprise complete words:

[1] Throughout the paper, the symbol <> indicates pause location.

3 <>*the amount of input of "r[ael]eal" [nativ] second language input [yh] they seem* <>.

The analysis of productivity, focussing on the nature of units or text spans produced between pauses, entails a careful and systematic description of the location of pauses appropriate to the on-line data we are attempting to describe. In the following section, we present such a framework for this categorisation, tentatively proposing a number of improvements to the approaches traditionally used in psycholinguistic research.

4. APPROACHES TO THE DEFINITION OF PAUSE LOCATION

There is a long history of psycholinguistic research concerned with the identification of pausing or non-fluency in spoken language production (Henderson et al 1966, Goldman-Eisler 1968, Butterworth 1980, Beattie 1983, and so on). Such work was principally concerned with the correspondence between location of non-fluency (associated with planning) and the structural composition of the message, be it grammatical or phonological. The widely reported claim is that non-fluencies commonly occur at clause and sentence boundaries. Goldman-Eisler (1972), for example, found that the longest pauses appeared between sentences or before coordinated clauses, and the shortest within clauses.

It is not only the major constituent (clause) boundaries, however, which attract pauses. On the basis of a study of impressionistic non-fluency locations in spontaneous speech, Garman (1990) suggests that certain clause-internal locations (for example, following an utterance-initial connective, and before an adverbial constituent) are more susceptible to non-fluencies than others. Other positions, too, within the noun phrase attract pausing. Echoing the claim of Maclay & Osgood (1959) concerning pausing before content words, Garman discusses the association between pausing and lexical selection.

Other investigations of pause behaviour, this time obtained from reading aloud and parsing tasks, focus on the (non-)correspondence between pausing and grammatical structure. Grosjean & Deschamps (1975), Grosjean, Grosjean & Lane (1979) and Gee & Grosjean (1983) use hierarchical 'performance structures' based on length of pause to cluster constituents in the text string. It is interesting that mismatches occur between grammatical and performance structures, for example, when pausing intervenes in a noun phrase unit such as *the strange* <> *young man*.

The analysis of pause location in spoken language production needs, of course, to look beyond explanations of a grammatical nature. The correspondence between units of production and prosodic structure is extensively reported in such sources as Halliday (1967), Boomer (1965), Chafe (1980), Brown and Yule (1983). Foster et al's (2000) overview of units of spoken language production, mentioned earlier, lists a number of such intonationally defined units as the tone unit, idea unit, and utterance, in contrast to units defined principally on syntactic or semantic criteria.

For the purpose of describing on-line written language production, however, we should refer to previous studies of pausing in writing (e.g. Matsuhashi 1981, 1982, 1987, Warren 1996, Janssen et al 1996, Schilperoord 1996, Stromqvist & Ahlsen 1998), which have tended to identify units of production solely on the basis of grammatical notions, such as sentence, clause, phrase and word. An exception to this, however, is the work of Sanders et al (1996) and Schilperoord & Sanders (1999), who consider the correspondence between pausing and text structure, looking at the hierarchical arrangements of segments connected by clause relations such as claim-argument, problem-solution and sequence. This direction opens up the interesting possibility of exploring the association between text production and topic development (continuity and discontinuity). The discourse orientation of this approach is one source of inspiration for the framework we develop for our own study. We return to this issue later.

Another goal for our study is to develop a framework which more sensitively accounts for distinctions at the level of 'word'. Following earlier critical observations (e.g. Spelman Miller 1997) concerning the lack of refinement of existing categorisations, especially in the definition of word-level locations, our framework seeks to address important distinctions, for example, between words of different classes (determiner, noun, disjunct, conjunction, and so on).

A third area of interest is to develop an approach which is more suited to the on-line nature of the data in our study. This entails the representation of location in terms of potential as well as actual (produced) units of language. Again, this issue is illustrated further below.

5. GENERAL FEATURES OF THE FRAMEWORK

As an essential step in identifying the units of written text production, pauses are categorised according to their locations within the stretch of text. The definition of these locations depends on the identification of structural elements within the text, and for this we draw on the concept of rank scale (Thompson, 1996; Halliday 1985, 1994) which distinguishes the following analytically useful units of text: morpheme, word, group, and clause. The basis for this rank scale is that 'units at each rank can be made up only of units from the rank below', and that 'every element is accounted for at each rank' (Thompson 1996: 21).

Such a classification corresponds to functionally distinct elements within the stretch of text, with the clause the highest rank or level, and the morpheme the lowest. In the case of our own data, however, we also include the level of sentence, which is useful in describing stretches of text signalled through the presence of terminal punctuation marks (full stop, exclamation mark, question mark, semicolon). In sum, the basic units used for the definition of pause location in our scheme are: character (morpheme), word, intermediate constituent (group), clause and sentence.

The identification of pause location is made in terms of the *preceding* structural element or elements. The decision to define position within the flow of the

text in this way allows us to arrive at a simple and consistent descriptive framework, which avoids the problems of multiplicity mentioned in Garman (1990):

> One way to document the occurrence of non-fluencies in the data is to examine their preceding and following contexts. These may be quite diverse, and difficult to determine in a unique fashion. (...) Tabulating all 136 non-fluencies in this way appears to require something like twenty-one preceding and twelve following context categories, or a total of 252 combinations. (1990: 120)

As we mentioned briefly above, a further feature of our approach to location arises from the on-line nature of our data. Given that location is to be defined in terms of a text string which is emergent and open to modification rather than static and complete, categories of structure selected to identify location can only reflect the status of the unit at a particular point in the construction of the text. Location, therefore, is more appropriately defined in terms of potential rather than definitive structural units. For example, in the case of 4a, the pause may appear to follow the noun phrase, *research*:

4a *research<>*

However, as the text string emerges this unit may be modified (subject to context), for example to 4b or 4c, in which case the pause now appears to be word-internal.

4b *research<>ers*
4c *research<>ed.*

The way we have chosen to reflect this is to categorise each pause not retrospectively on the basis of the completed string, but according to the potential structure at the specific point of its occurrence. Such an on-line analysis attempts to acknowledge the range of options available as the writer produces the text, and avoids retrospective categorisation of the product.

To accommodate this notion of potentiality, the locations are referred to as *potential completion points*. This term reflects the idea that the development of the text is not predetermined but locally 'negotiated' between writer and text, and at any one point the text could develop in a number of different directions. In other words, it allows for the fact that the emergent status of a text unit may be defined in a particular way, while recognising that this status is not fixed and may alter as the text subsequently develops.

Finally, the choice of the term '*completion* point' should not be taken to imply the function of the pause in completing (or producing) the relevant text unit. As we have mentioned above, the categorisation refers simply to the position of the pause within the flow of the text. It may be that the occurrence of a space before the pause will help to confirm its status as word (or group) versus non-word. For example in the following extract,

5 *This_is_<1>because_<2>,<3>children<4>_receive_enough_<5>e<6>xposure_to _<7>L2<8>_*

We can fairly safely assume that the spaces (marked as _) indicate a boundary after discrete words (*is, because, enough, to*). In the case of <4>, however, where the pause is not preceded by a space, it is not possible to know at the point of the pause whether *children* was the intended form, or whether the target was *children's*, or possibly even was to be modified to *childish, childcare,* and so on. The possibility of future modifications to the item *children*, however, does not alter the status of the pause at this point as following a noun phrase, or nominal group.

6. GRAMMATICAL LOCATION CATEGORIES

The categories used in the analysis of pause location are as follows:

– sentence completion point (SCP)
– clause completion point (CCP)
– intermediate constituent (or group) completion point (ICP)
– word completion point (WCP)
– character completion point (XCP)

The categorisation is designed as a working model to capture the main locations of interest in our data, and cannot claim to address all possible occurrences. The procedure for analysing is summarised on a flow-chart (see Figure 2) as a series of decisions with respect to potential completion points based on definitions of the relevant text units. The chart begins with the completion point following the largest text unit, the sentence, and proceeds through clause, intermediate constituent, word to character and morpheme levels.

Appendix A presents a summary of the main contents of the categories SCP, CCP, ICP, WCP and XCP which are defined and illustrated below.

6.1. SCP (Sentence completion point)

As we mentioned above, a decision was made to categorise pauses at sentence completion points where an explicit terminal punctuation mark is used (full stop, exclamation mark, question mark, semi colon and, in some cases, colon), regardless of subsequent modifications to the text. Carriage returns for paragraph settings were not indicated separately, but coincide with sentence boundaries.

6.2. CCP (Clause completion point)

This refers to the location of a pause at a potential or emergent clause boundary; that is, where the text unit is structurally and semantically acceptable as a clause,

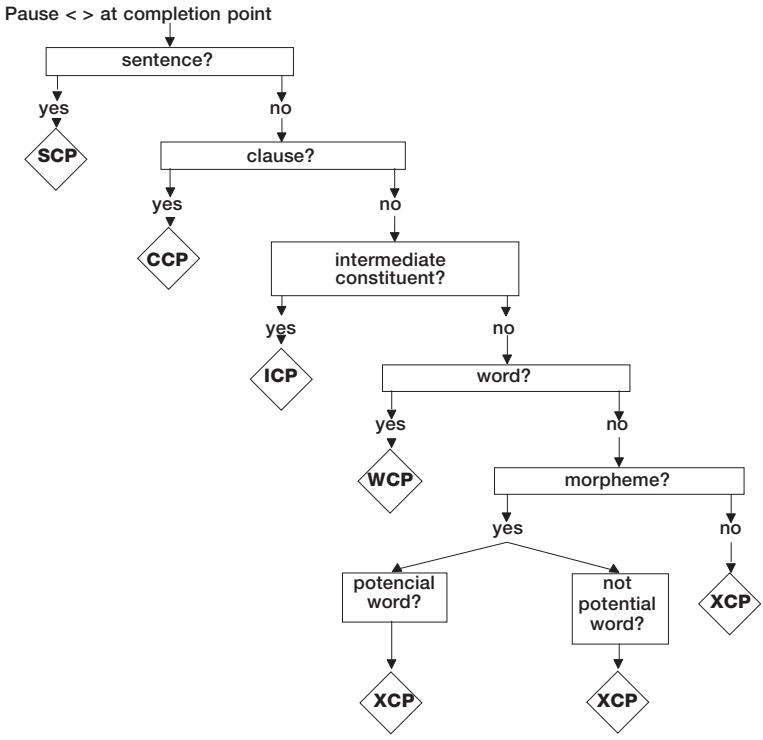

Figure 2. Identification Procedure

but is not bounded by terminal punctuation. (In some cases, but not all, a comma is used to signal the clause boundary.) The category includes clauses containing a finite verb independent or subordinate), non-finite verb forms (-ing, -ed), verbless clauses (eg. *whether right or wrong,<>),* regardless of subsequent modifications. This leads to the categorisation of the pause location in example 6 below as a CCP, since the clause is both structurally and semantically acceptable as a clause at the point of pausing, without the addition of the prepositional phrase, *in the area*:

6 *...has been the focus of interest in many studies <> in the area*

Context is clearly important in determining whether or not a potential clause completion point has been reached, and illustrations will be given below of cases where the analysis is not always straightforward.

6.3. ICP (Intermediate constituent completion point)

This is the largest and most complex category and includes both grammatically and discoursally significant components. These are outlined below, taking firstly the issue of grammatical constituency within the clause.

The concept of an intermediate level constituent between clause and word draws on the notion of the grammatical unit which fills a functional slot within the clause, but does not complete the clause. The functions within the clause which are particularly important for our analysis are those of Subject and Verb or Predicator (Thompson, 1996: 18-19), and less so, Object, Complement and Adjunct, again when these do not coincide with potential clause completion. The concept of intermediate constituent overlaps with the category of group as discussed by Lock (1996) and Thompson (1996). Of particular importance in our analysis are the constituents which fill the functional slots of Subject and Predicator. These tend to coincide, respectively, with nominal and verbal groups, which may consist of single or multiple word groups.

The nominal group, for example, may take a range of forms, such as the units preceding the pauses in:

– *research <> has*
– *the tests used in the study <> were*
– *the relation between age and SLA <> is*
– *one of the explanations for this <> is*
– *the research that has addressed the age issue <> is...*

In each of these cases, the text unit preceding the pause fills the functional slot of Subject, and the nominal group constitutes an intermediate constituent. Postmodification of the nominal groups here does not affect their status as Subjects. Postmodification, as Thompson (1996: 184) outlines, 'has the primary function of add[ing] specificity to the nominal group', and typically involves an embedded phrase or clause, which may be finite (eg. a defining relative clause or projected clause, such as 'a feeling that...') or non-finite. In 8a and 8b, for example, both nominals fill the same Subject slot:

8a *the research <> is*
8b *the research that has been done <> focuses...*

Although our primary concern is with Subject slot fillers with respect to nominal groups, the case of nominal groups which are in Complement or Object slot should also be mentioned. It is of course possible (although relatively rare) that such slots are fronted in the clause, in which case the text units would be considered intermediate constituents. In post-verbal position, nominal groups (Complement/Objects) frequently serve to complete the clause, in which case, the relevant text unit for our analysis becomes the clause. An illustration of this is given in extract 9 below, where both pause <1> and pause <2> are categorised as CCP:

9 *learners can improve aptitude <1>through greater practice<2>*

In some cases, postmodification will require context-sensitive analysis. It may be argued that the pause in example 10 can be considered at a CCP.

10 *there are four factors <> to consider*

However, this raises the issue of the centrality of the postmodifier to the nominal group. Decisions need to be made in such cases as to whether the nominal group can stand without the postmodifier, as in example 11 below, or whether, as in example 12, the nominal group is not complete at the point of the pause:

11 *which affect the process <> of*

12 *learners have an interest<> in language*

To sum up so far, one of the main categories of interest in our analysis is that of items filling the functional slot of Subject, and these coincide for the large part with nominal groups, including those postmodified by such elements as prepositional phrases and relative clauses. Another important category which we consider as intermediate constituents is that of the *verbal group*, which we define as main verb and other auxiliary verbs. This can be illustrated by forms such as the following from our data:

13a (*he*) *said* <>
13b (*some*) *believe* <>
13c (*aptitude*) *has been criticised* <>
13d (*which*) *may be* <>

In the cases of pauses immediately after 'be' and 'have', it is necessary to use context wherever possible to distinguish auxiliary from main verb use. Often the default assumption has to be that the main verb could stand in that position, and therefore the text unit is coded as an intermediate constituent. Where a pause occurs between auxiliary and verb, this is considered to be a group-internal position, and is categorised as a WCP (see below).

Phrasal verbs, that is a verb + particle structure, present an interesting case for analysis, since the process is being expressed not only by the verbal group proper, but also by an additional constituent, the particle. In cases where the pause occurs after the phrasal verb, as in example 14,

14 *find out* <>

the whole text unit is taken as an intermediate constituent. In many cases, the pause will occur after the verb only, as in 15:

15 *found* <> *out that*

and here the verbal element alone will constitute a potential intermediate constituent, particularly since the particle structure is only one of a number of continuations of the string.

The verbal group complex includes structures such as verb + 'to'-infinitive , or '-ing' form (Thompson 1996: 191). Where a pause occurs after the first or second verbal element, the category ICP is appropriate, but we identify the location after the 'to', for example in 16,

16 *wanted to* <> *investigate*

as falling within the verbal group, and hence at a WCP.

Other interesting cases which arise in the categorisation of intermediate constituents include prepositional phrases, and complex adjectival groups such as *capable of, happy with, unlikely to*. In these cases, as before, we have to consider the centrality of the postmodifier to the meaning of the head, and the potential completeness of the unit at the point of the pause.

We began our discussion of the category of intermediate constituent by suggesting that grammatical constituency was not the only issue to be addressed. We wish also to account for a number of discoursally significant elements, those of *disjuncts*, *conjuncts*, and *conjunctions*, which may be thought of as sitting outside the nuclear structure of the clause, but nonetheless carry a particular importance in organising the discourse. Examples of these are:

17a *in particular,* <>
17b *however,* <>
17c *he said that*<>
17d *although* <>
17e *in other words,* <>
17f *whereas* <>.

In some cases, but clearly not all, these may be set apart from the rest of the clause by the use of a comma.

6.4. WCP (Word completion point)

This category refers to cases where a pause falls at a recognisable word boundary (emergent or potential), regardless of subsequent modifications, and when the text unit does not constitute an intermediate constituent, as discussed above. In other words, this category will include within-group locations such as:

– determiner <> noun
– auxiliary <> main verb
– preposition <> noun phrase
– after conjunctions within nominal or verbal group complexes (eg. *age and* <> *motivation*)

and other locations within groups, the most common of which include:

– determiner + noun <> clause/phrase (eg. *interest* <> *in learning*)
– adjective <> clause/phrase (eg. *capable* <> *of learning*)
– after a relative pronoun (eg. *the research that* <> *has been*)
– within verbal group complexes (eg. *try to* <> *learn*).

Another category to be included is that of adjuncts which are integrated in the group structure, such as:

18a *they can now* <> *begin*
18b *it is very* <> *important to*
18c *they really* <> *want to*
18d *the learner may also* <> *have*
18e *motivation is related closely* <> *to success*

where the adjuncts are not realised by adverbial groups or prepositional phrases (which would constitute intermediate constituents in our analysis), and where they do not occur at a potential clause completion point.

6.5. XCP (Character completion point)

This refers to the position after a recognisable character or sequence of characters (and since the data is typed, these are executed rather than potential characters), which do not constitute a word. These word-internal locations may or may not coincide with morpheme boundaries. In 19 below, for example,

19 *e<>xposure*

the pause does not occur at a morpheme boundary. Some morphemes may also be potential words, as in the case of *in* in example 20:

20 in_in<>formal_contexts

Here, grammatical context has to be used to identify this pause at a word-internal or character-level location, rather than at a viable word boundary.

7. FROM GRAMMAR TO DISCOURSE: THE FRAMING DEVICE

Our earlier discussions identified as a key area for investigation the association between the processes of text production and the substance of the text produced, the 'critical juncture between planning and translating' (Witte & Cherry 1986: 143). While the focus of the previous section has been on characterising the language produced in terms of grammatical status, the purpose of this next section is to consider textual output from a more discourse-oriented perspective. There has

been little attention so far in pausological studies of writing to issues of discourse, and yet the association between text structure and the processes involved in handling topic appear well worth pursuing.

With this in mind, we propose a scheme to describe units of language produced on-line from the perspective of certain discourse functions associated with the notion of topic. Put simply, as an extension to the syntactic categorisation of units in the analysis of pause location, we propose an additional category called the *framing device*, which allows the interpretation of certain text units in terms of their role in introducing and developing topic in the discourse. The elements we focus on occur in the intermediate constituent category (nominal groups in subject or adjunct position, and also conjuncts and disjuncts) and certain clause-level elements which may be seen as framing or setting up the rest of the message.

The choice of the term *framing* device is influenced by the use by Witte & Cherry (1986), following Bracewell et al (1982), of the concept of framing strategy. By analogy with the notion of topic-related processes in conversation (Bracewell et al 1982: 148), this refers in the case of written language to processes by which the writer makes choices about topicalisation in clauses and the establishment of topical relations across clauses (Witte & Cherry 1986: 127).

The notion of topic is one which is referred to widely in the literature, and defined in a bewildering variety of ways. Traditionally, approaches to topic have focused more on the notion of sentence rather than discourse topic, and in so doing, have employed a range of defining properties to identify topic. In a recent review, Goutsos (1997: 5) lists the following perspectives on sentence topic: structural, presentational, logical, informational and pragmatic. Each uses different clusters of properties in the definition of topic. A structural perspective, for example, specifies grammatical conditions for topic, whereas a presentational approach relates topic to position in the clause or sentence (leftmost constituent) or to staging, that is, the presentation of clause or sentence elements in relation to a point of departure (Brown & Yule 1983). A further view relates topic to the notion of what the sentence is about. This aboutness view or logical perspective (Reinhart 1981) may provide a link to the analysis of topic at a discourse level. Other approaches connect topic with information structure (that is, the status of elements as given and new) or with pragmatic notions of prominence. The range of perspectives offered in the literature is clearly considerable, and some selection of approach is necessary for the purposes of our study.

The first decision is to focus on topic at the local (sentence) rather than discourse level. This is the level which much of the psycholinguistic (production processes) literature in general and pausological research in particular has been concerned with. Such interest in sentence-level topic allows connections to be made to other complementary sentence-level topic analyses of writing, such as Lautamatti (1987) and Witte (1983)'s use of topical structure analysis, which explores relations between initial sentence element, grammatical (or mood) subject and topical subject (or sentence topic).

Since our analysis is concerned with characterising the units of text produced on-line, there is a clear justification for selecting a local-level approach to the

description of topic. From our local perspective in analysing text at the point of production, we are not in a position to know the consequences or effect in global, discourse terms of the selection of certain units in the text string. Indeed, an element may not survive phases of revision to appear in the final text product, or its function may not be clear at the point at which it is produced. Equally, the function of an item in the text string in terms of its contribution to the development of the discourse topic may not become clear until the whole text is produced. Since our study aims to describe text units as they emerge at the point of production (and without knowledge of their ultimate effect on the overall text structure), we set the scope of our consideration of topic, therefore, at the level of sentence/clause.

The second decision concerns the perspective taken in defining our topic-related notion, the framing device. As we have noted above, a number of different defining properties of sentence topic have been proposed in the literature. Given the grammatical approach taken so far in our study (and in other pausological studies) to the identification of pause location, we are drawn to notions of structure (the grammatical manifestation of topic within the clause or clause complex) and presentation (in particular, initial clause/sentence position) for the definition of the notion of framing device. Both of these perspectives sit well with the Hallidayan view of *theme*, that is, of the first ideational element within the clause as the 'point of departure for the message' (Halliday, 1985: 38), which inspires the categories of framing device we propose.

Notions of grammatical form and sentence position are in many ways complementary. In English, of course, word order preferences or constraints will dictate that theme and subject often coincide, with the result that subjecthood and initial clause/sentence position are often in danger of being conflated. They are, of course, separate issues: adjuncts, complements and objects may be thematic (that is, they are marked as fronted themes), constituting the starting point of the message but not the grammatical subject. Grammatical subjecthood does not entail ideational meaning. In the case of light or 'empty' themes in clause initial position (ie. *it* and existential *there*), for example, the *it* and *there* may be in subject slot but do not carry ideational meaning.

In general terms, for the purposes of our study, we draw on both structural and presentational notions of topic to describe elements in our data. The categories of framing device we establish draw on terminology, especially from the Hallidayan perspective, to define topic-related functions in a rather general and eclectic manner. The terms we propose are presented and illustrated in the following section.

7.1. General definition of framing devices

Our notion of framing device is defined quite simply as an element or structure (single word, phrase or clause) which serves to establish the starting point of the message at the clause/sentence level. This may be in one of a number of ways, either in constituting the topic itself, or in preparing the scene for the introduction of the topic, for example through the use of a discourse marker such as *however*,

1.	Subject theme
2.	Adjunct theme/complement theme
3.	Non-experiential theme
4.	Empty theme (*it*, *what* and existential *there*)
5.	Thematised structure (eg finite/nonfinite clauses)

Figure 3. Categories of framing device

empty theme such as *it is (stated) that*, or initial clause structure such as *By doing this*. In conception our approach bears some similarities to that proposed by Goutsos (1997) who considers topic-related strategies visible in the text product which are associated with topic framing, topic introduction, topic closure and topic continuation.

The taxonomy we present (see Figure 3) consists of five types of framing device which fulfil these topic-related functions. They are robust, low-inference categories which can be applied to the existing pause location categories (ICP and CCP) described above. In the section below we briefly define these categories.

7.2. Subject theme

This category refers to elements which are both grammatical subject and initial sentence constituent. It contains both full nominal groups (including those which are postmodified) and proforms. This is the most frequently occurring category in our analysis. In the examples below the relevant exponent is illustrated (underlined) in context.

21 This hypothesis might be attributed to the claim of lateralisation.

22 This would be further discussed in terms of the claim of 'the Affective Filter Hypothesis'.

7.3. Adjunct theme/complement theme

Sentence-initial adverbials in particular serve the important function of preparing for the introduction of a new topic. Examples of this are as follows:

23 Around puberty, human beings will face with lateralisation of the brain.

24 Among individual factors, those which are widely recognised by most scholars are: age, aptitude, attitude, motivation and personality.

25 With reference to Ellis (1994), he illustrates these differences with the aid of a framework.

7.4. Non-experiential theme

In Halliday's model of theme, the experiential function contrasts with (or indeed complements) the textual and interpersonal functions in text. In identifying the non-experiential theme, then, as a category in our analysis, we are drawing on concepts of textual structuring and interpersonal (evaluative) comment to define the elements which appear in sentence initial position, but do not constitute part of the ideational content of the message. In Goutsos's model, such elements are frequently used to signal topic shift, continuation or closure.

This category includes disjuncts (sentence adverbials) and conjunctions which in Hallidayan terms serve a function outside the ideational content of the message. In other words, although sentence initial, they do not 'exhaust the thematic potential of the clause' (Halliday 1994: 52, cited in Thompson 1996: 134). Thompson (1996) describes this non-experiential function in the following terms: 'they indicate the location of the starting-point in the text's semantic space without in themselves constituting the starting-point' (1996: 136). In the case of conjunctions, their purpose is to signal how one clause relates to the larger clause complex. Sentence adverbials (conjunctive adverbials and modal adjuncts) provide some kind of structuring or evaluative function on the content of the message, and in this sense are outside the ideational content itself (Lock 1996: 229-231). In written text, such forms are often separated from the rest of the clause by a comma, in a way which marks them out as non-integral to the information of the clause. It is of course true that (with the exception of conjunctions whose position is fixed) these items may occur in non-thematic positions within the sentence, and that the choice by the author to thematise is therefore of interest.

The range of structures to be included in the category of sentence adverbial is vast, including adverbs, prepositional phrases, infinitive clauses, *-ing* and *-ed* participle clauses and finite verb clauses (Leech & Svartik 1975: 201). The category includes, then, items which both convey authorial comment on the content (eg. *of course, theoretically*, etc.) and perform a connective function in structuring the content of the message (eg. *therefore, furthermore*, etc). Examples in context from our data are given below:

26 In addition, Long (1990) provides evidence to suggest that acquisition of a native-like accent is not possible after the sixth year of life.

27 To start with, in an attempt to present a theoretical view of motivation, Skehan put forward four hypotheses.

28 Moving on to some more general factors that influence the performance and the acquisition of the learners of a SL, we cannot avoid referring to motivation.

7.5. Empty theme (*It*, *what* and existential *there* structures)

In the case of existential *there*, which occurs in subject position but does not in itself fulfil the experiential function of theme, the decision has been made, following Thompson (1996: 138) but contrary to Halliday (1994: 44) to include the existential process (realised by the verb 'be') in the scope of the theme. This allows us to account for the existential process as theme or starting point for the rest of the clause. Examples of this in our data include:

29 there are three major dimensions of the variable personality.

30 There are debates on the methodology of experiments carried out.

A second structure which we include in this category is the pseudo-cleft structure or thematic equative, as in:

31 What is needed in this area of research is some longitudinal studies.

A third case of multiple constituent theme which we handle within this category is the construction including *it,* where this appears in first slot theme position, but functions as an empty place-holder (dummy subject), with the real subject being the second clause. Examples of this from our data include 'predicated theme' (cleft) constructions, as in:

32 it is that young learners acquire a language more easily.

and 'thematised comment', where the whole structure of the *it is X that* clause sets up the content of the following clause. The comment of the *it is X* clause often consists of a value comment on the following content, as in:

33 It is well known now that individuals acquire language in different ways.

34 It may be the case that (...) in the beginning group one students might learn faster.

35 It is beyond doubt that everyone apart from some exceptions can learn a language.

In these cases, the decision has been to include the *that* within the thematic unit since this frequently seems to be the boundary marker of the clause.

7.6. Thematised structure

Finally, this category includes a variety of structures, both clausal (finite and non-finite) and phrasal, which are fronted in the sentence. These differ from

examples of non-experiential theme, since they are integral to the content of the text, rather than being structuring or authorial comment on the content. Examples of such structures include:

36 Because of a number of individual differences in the same age group, there should be different processes of L2 learning and acquisition.

37 If the teacher knows that not all students have the same learning strategies, he/she will vary the material.

38 Since I was a child, his big dream was I to become an English teacher just like him.

To summarise the goal of the paper so far, we have sought to present and illustrate a framework for the analysis of on-line written text production as a means of characterising various units of production. Underpinning such an enterprise has been the development of a scheme for the identification and interpretation of pause location, which we discuss in some detail. In a number of important ways, the proposed framework extends and improves other schemes used in pausological research. Firstly, it offers a more elaborate scheme for the identification of elements, particularly at the word level. Secondly, it represents location in terms of potential completion points. Such an approach, we argue, offers a more appropriate means of accommodating the emergent and often temporary status of text elements as they are produced on-line. Thirdly, the notion of location is extended to incorporate a discourse-sensitive category, the framing device, which allows us to interpret certain text units in terms of topic-related functions.

In the final section which follows, we briefly refer to the application of these categories to the study of writing processes and textual output. This will allow us to speculate on the type of insights such an analysis might bring to the investigation of writing behaviour.

8. APPLICATION AND EVALUATION

The intention behind presenting this detailed framework is to encourage the application of such a text analytical approach to different contexts of written language production. In our own study (Spelman Miller 1999, 2000a, 2000b) we investigated the writing processes of a number of L1 and L2 writers producing academic essays under timed, simulated examination conditions. The two tasks given to the subjects varied in rhetorical demand but were in general terms based on the same topic area. The broad aim of the study was to compare performances of subjects on the two tasks in terms of the pause-related phenomena outlined in section 3 above.

Measures of the main dependent variables, pause duration, pause frequency and rate, productivity and rate of production, revealed predicted differences be-

tween the L1 and L2 writers. The L2 writers produced longer pauses at all grammatical locations, and especially at word-internal and intermediate constituent locations, suggesting increased attention to lower-level processing concerns. Productivity and rate of production were also significantly lower in the case of the L2 writers. On all measures, however, the writers behaved in very similar ways under the two task conditions, suggesting a lack of perception of, or ability to respond to, the differing rhetorical demands of the two tasks.

Of particular interest to our present discussion is the application of the framing device unit to data in our study. As we have outlined above, this exploratory unit has been proposed as a means of capturing structures in the text string associated with the establishment, maintenance and development of the topic. The selection of exponents of this framing device category is far from definitive, but includes a number of key units such as subject theme, non-experiential theme (including disjuncts and conjunctions), and thematised structures, that is, fronted clausal and phrasal structures. Indeed, our initial observations reveal that these three are by far the most frequently occurring framing device categories, with subject theme being the most prominent category.

Our analyses so far seem to support the notion of the framing device as a potentially interesting means of interpreting the data string produced during composition. In general terms, we found an interesting coincidence of framing device and pausing: approximately one third of framing device locations coincided with the boundary of a unit of production (i.e. a pause). This suggests a potentially powerful role for these framing devices in the writer's management of the formulation process.

In more specific terms, the analysis of framing devices appears to open a window on the subtle processing preferences of individual writers when producing text. Investigation of pause duration at framing device locations reveals significant differences between the subject groups, with the L2 writers pausing for considerably longer at these locations than the L1 writers.

Finer-grained case study analyses of two individual L2 and one L1 writer confirm this general observation, but also point to intriguing and complex differences between individuals. For the two L2 writers, but not for the L1 writer, the framing device location appears to be a natural location for pausing, although the nature of the pause behaviour differs. In the case of one L2 writer, subject themes are particularly susceptible to pausing, although pauses at these locations tend not to be lengthy. For the other L2 writer, pausing is less frequent at these framing device locations but when it does occur pauses tend to be very long. Finally, in the case of the L1 writer, extensive pausing occurs almost exclusively at clause and sentence locations and not at framing device locations.

These observations, while being tentative and speculative, based on a small sample of writers, allow us to glimpse the complexity of the writing process viewed from a cognitive-textual perspective. Moreover, in general terms, they highlight the need for the careful definition and application of units of analysis, such as those presented in this paper, and encourage the refinement and extension of the type of framework we have offered here for the description of units of written text production.

WORKS CITED

BEATTIE, G.W. *Talk: An Analysis of Speech and Nonverbal Behaviour in Conversation.* Milton Keynes: Open U, 1983.

BONIN, P. and M. FAYOL. "Producing Isolated Words from Pictures, from Orally and Visually Presented Words: An On-line Study of Naming and Writing." *Current Research in Writing: Theories, Models and Methodology.* Ed. G. Rijlaarsdam, H. van den Bergh & M. Couzijn. Amsterdam: Amsterdam UP, 1996. 145-158.

BOOMER, D.S. "Hesitation and Grammatical Encoding." *Language and Speech* 8 (1965): 148-158.

BRACEWELL, R., C. FREDERIKSEN and J. FREDERIKSEN. "Cognitive Processes in Composing and Comprehending Discourse." *Educational Psychologist* 17 (1982): 146-164.

BROWN, G. and G. YULE. *Discourse Analysis*. Cambridge: Cambridge UP, 1983.

BUTTERWORTH, B. "Evidence from Pauses in Speech." *Language Production: Vol. 1. Speech and Talk.* Ed. B. Butterworth. London: Academic, 1980. 155-176.

CHAFE, W. "The Deployment of Consciousness in the Production of a Narrative." *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Ed. W. Chafe. Norwood, NJ: Ablex, 1980.

COOPER, M. and M. HOLZMAN. "Talking about Protocols." *College Composition and Communication* 34 (1983): 284-293.

ERICSSON, K. and H. SIMON. *Protocol Analysis: Verbal Reports as Data.* Cambridge: MIT, 1984.

FOSTER, P., TONKYN, A. and G. WIGGLESWORTH. "Measuring Spoken Language: A Unit for All Reasons." *Applied Linguistics* 21 (2000): 354-375.

GARMAN, M. *Psycholinguistics*. Cambridge: Cambridge UP, 1990.

GEE, J. and F. GROSJEAN. "Performance Structures: A Psycholinguistic and Linguistic Appraisal." *Cognitive Psychology* 15 (1983): 411-458.

GOLDMAN-EISLER, F. *Psycholinguistics: Experiments in Spontaneous Speech.* London: Academic, 1968.

—— "Pauses, Clauses, Sentences." *Language and Speech* 15 (1972): 103-113.

GOULD, J. "How Experts Dictate." *Journal of Experimental Psychology: Human Perception and Performance* 4 (1978): 648-661.

—— "Experiments on Composing Letters: Some Facts, Some Myths, and Some Observations." *Cognitive Processes in Writing.* Ed. L. Gregg and E. Steinberg. Hillsdale, NJ: L. Erlbaum, 1980. 97-127.

GOUTSOS, D. *Modeling Discourse Topic: Sequential Relations and Strategies in Expository Text.* Norwood, NJ: Ablex, 1997.

GREENE, S. and L. HIGGINS. "'Once upon a Time': The Use of Retrospective Accounts in Building Theory in Composition." *Speaking about Writing*. Ed. P. Smagorinsky. London: Sage, 1994. 115-140.

GROSJEAN, F. and A. DESCHAMPS. "Analyse contrastive des variables temporelles de l'anglais et du francais: vitesse de parole et variables composantes, phenomenes d'hesitation." *Phonetica* 31 (1975): 144-184.

GROSJEAN, F., GROSJEAN, L. and H. LANE. "The Patterns of Silence: Performance Structures in Sentence Production." *Cognitive Psychology* 11 (1979): 58-81.

HALLIDAY, M.A.K. "Notes on Transitivity and Theme in English." Parts 1-3. *Journal of Linguistics* 3 (1967): 37-87; 3 (1967): 199-244; 4 (1968): 179-215.

—— *An Introduction to Functional Grammar.* London: Edward Arnold, 1985.

—— "The Construction of Knowledge and Value in the Grammar of Scientific Discourse, with Reference to Charles Darwin's *The Origin of Species." Advances in Written Text Analysis.* Ed. M. Coulthard. London: Routledge, 1994.136-156.

HENDERSON, A., F. GOLDMAN-EISLER and A. SKARBEK. "Sequential Temporal Patterns in Spontaneous Speech." *Language and Speech* 9 (1966): 207-216.

JANSSEN, D., L. VAN WAES, and H. VAN DEN BERGH. "Effects of Thinking Aloud on Writing Processes." *The Science of Writing.* Ed. C.M. Levy and S. Ransdell. Mahwah: L. Erlbaum, 1996. 233-250.

KOWAL, S. and D.C. O'CONNELL. "Writing as Language Behavior: Myths, Models and Methods." *Writing in Real Time: Modelling Production Processes.* Ed. A. Matsuhashi. Norwood: Ablex, 1987. 108-132.

LAUTAMATTI, L. "Observations on the Development of the Topic of Simplified Discourse." *Writing across Languages: Analysis of L2 Text*. Ed. U. Connor & R.B. Kaplan. Reading, MA: Addison-Wesley, 1987. 87-114.

LEECH, G. and J. SVARTIK. *A Communicative Grammar of English*. London: Longman, 1975.

LOCK, G. *Functional English Grammar*. Cambridge: Cambridge UP, 1996.

MACLAY, H. and C.E. OSGOOD. "Hesitation Phenomena in Spontaneous English Speech." *Word* 15 (1959): 19-44.

MATSUHASHI, A. "Pausing and Planning: The Tempo of Written Discourse Production." *Research in the Teaching of English* 15 (1981): 113-134.

—— "Explorations in the Real-time Production of Written Discourse." *What Writers Know: The Language, Process, and Structure of Written Discourse.* Ed. M. Nystrand. New York: Academic, 1982. 269-290.

—— *Writing in Real Time: Modelling Production Processes.* Norwood: Ablex, 1987.

REINHART, T. "Pragmatics and Linguistics: An Analysis of Sentence Topics." *Philosophica* 1 (1981): 53-94.

RUSSO, J.E., JOHNSON, E.J. and D.L. STEPHENS. "The Validity of Verbal Protocols." *Memory and Cognition* 17 (1989): 759-769.

SANDERS, T., JANSSEN, D., E. van der POOL, J. SCHILPEROORD and C. van WIJK. "Hierarchical Text Structure in Writing Products and Writing Processes." *Current Research in Writing: Theories, Models and Methodology.* Ed. G. Rijlaarsdam, H. van den Bergh & M. Couzijn. Amsterdam: Amsterdam UP, 1996. 473-492.

SCHILPEROORD, J. "The Distribution of Pause Time in Written Text Production." *Current Research in Writing: Theories, Models and Methodology.* Ed. G. Rijlaarsdam, H. van den Bergh & M. Couzijn. Amsterdam: Amsterdam UP, 1996. 21-35.

SCHILPEROORD, J. and T. SANDERS. "How Hierarchical Text Structure Affects Retrieval Processes: Implications of Pause and Text Analysis." *Knowing What to Write: Conceptual Processes in Text Production.* Ed. M. Torrance & D. Galbraith. Amsterdam: Amsterdam UP, 1999. 13-30.

SEVERINSON EKLUNDH K. and P. KOLLBERG. "A Computer Tool and Framework for Analyzing Online Revisions." *The Science of Writing.* Ed. C.M. Levy & S. Ransdell. Mahwah: L. Erlbaum, 1996. 163-188.

SMAGORINSKY, P. *Speaking about Writing.* London: Sage, 1994.

SPELMAN MILLER, K. "Investigating Writing On-line: An Illustration of Methodology and Analysis." *University of Reading Working Papers in Linguistics* 3 (1997): 323-340.

—— "Investigating Academic Writing On-line; A Study of Temporal Aspects of Text Production." *Writing in an Electronic Environment: Research with Language Learners.* Ed. M. Pennington. Houston, Texas: Athelstan, 1999.

—— "Academic Writers On-line: Investigating Pausing in the Production of Text." *Language Teaching Research* 4 (2000a): 123-148.

—— "Writing on-line: Temporal Features of First and Second Language Written Text Production." Diss. University of Reading, UK, (2000b).

STROMQVIST, S. and E. AHLSEN. *The Process of Writing: A Progress Report.* 1st Draft. Department of Linguistics and SSKKII Center for Cognitive Science, U of Goteborg, Sweden, 1998.

THOMPSON, G. *Introducing Functional Grammar*. London: Arnold, 1996.

WARREN, E. "The Significance of Pauses in Written Discourse: A Comparison of Native Speaker and Learner Writing." *Second Language Acquisition and Writing: A Multidisciplinary Approach*. Ed. A. Archibald & G. Jeffery. U of Southampton: 1997. 152-177.

WITTE, S. "Topical Structure and Revision: An Exploratory Study." *College Composition and Communication* 41 (1983): 313-341.

WITTE, S. and R. CHERRY "Writing Processes and Written Products in Composition Research." *Studying Writing: Linguistic Approaches.* Ed. C. Cooper & S. Greenbaum. Newbury Park, CA: Sage, 1986. 112-153.

APPENDIX A: IDENTIFICATION OF PAUSE LOCATIONS (MAIN CATEGORY CONTENTS)

Key: D = determiner; N = noun; Aux = auxiliary; MV = main verb; S = subject Prep = preposition; NP = noun phrase; Cl/Phr = clause/phrase; O/C = object/complement; rel pron = relative pronoun; Adj = adjective; (examples given in brackets).