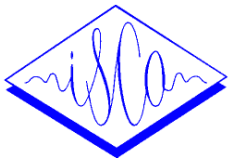
ISCA Archive
http://www.isca-speech.org/archive



ITRW on Disfluency in
Spontaneous Speech (DiSS'01)
Edinburgh, Scotland, UK
August 29-31, 2001

Disfluencies in Writing – are they Like in Speaking?

Åsa Wengelin

Department of Linguistics and Phonetics
Lund University, Lund
asa.wengelin@ling.lu.se

Abstract

This paper presents a study of disfluencies in written language production. Texts from ten university students are compared to data from people who almost never use writing, namely adult dyslexics and to texts from people who communicate in writing under real-time constraints every day, namely deaf whose main use of writing is text telephone conversations. This paper investigates which types of disfluencies occur in writing, where they occur and their durations. Further, this paper investigates how different text types and the specific characteristics of deaf and dyslexic writers influence the distribution of disfluencies. The results are discussed in relation to earlier work on disfluencies in speaking.

1. Introduction

While spoken communication is typically processed under strict on-line constraints where sender and receiver are located at the same point in time and often also in place, written language has until quite recently mainly been associated with communication where sender and receiver are located at different points in time and place. This is probably one reason why real-time processing phenomena such as pauses, hesitations, repetitions and repairs have mainly been associated with the spontaneous production of spoken language. Another reason is that until quite recently there have been no tools for 'recording' writing and therefore only the final products have been available while recordings of spoken language are always 'on-line'.

However, the use of computers has meant substantial changes for the study of writing. First, they have led to new and more on-line like uses of written communication, like email and chat. These are not studied in this paper. Second, repairs can now be recorded and pauses can be measured automatically by means of on-line logging. This paper presents studies made by means of automatic corpus analyses of on-line recordings of writing.

2. Written language production

2.1. On-line studies of writing

Before 1980 when Gregg and Steinberg [1] published their pioneering book *Cognitive processes in writing*, very little research had been done on writing as a cognitive process. Most earlier studies of writing had addressed pedagogical issues, but with Hayes and Flower's model [2], published in the book mentioned above, the interest started to focus more on the cognitive process of writing, and a new paradigm which attended to the writer's strategies to reach an acceptable solution to a rhetorical problem was developed. Since then there has been an immense increase in studies of written production, mostly by means of think aloud protocols, planning notes , retrospective reports and manual measuring of pause durations. The development of automatic on-line logging of writing, by means of computers has meant substantial changes for the study of written language production. However, there are still few on-line studies of writing around, and many insights could be gained from studies of spoken language.

2.2. Disfluencies and production conditions

In studies of spoken language disfluencies (df:s) have been studied from several different views. Many have been psycholinguistically orientated and in these df:s are often thought to be indicative of the mental processes underlying speech generation while others have focussed on how df:s are processed by the listener (e.g. [3]). A common view of disfluencies is that they indicate problems in the planning of utterances. However, it has also been argued that most disfluencies in spoken language production are not mistakes but signals for coordinating the speakers with their addressees (e.g. [4]), and, that both intraindividual and interindividual perspectives have to be taken into account (e.g. [5]). Most of these studies have been concerned with face to face dialogues. However, the typical writing situation is still the production of a monologue with no addressee present. The sender has plenty of time to plan, encode and revise the message before sending it. Therefore, the interactive function of dislfuencies can hardly be dominant in writing. There are also studies of disfluencies in spoken monologues. In [6] for example, it is shown that there is a system in how speakers produce filled pauses and that these often carry information about larger-scale topical units. Similar patterns could be expected to occur also in writing.

3. Methods

3.1. Subjects and Data Collection

The data used for the current analyses relate to ten Swedish university students, eleven Swedish adults with reading and writing difficulties, and nine congenitally deaf subjects. All subjects were all well acquainted with writing on a computer. The first language of the deaf writers is Swedish sign language, why, their writing my be considered as L2 writing. The data collection was made within the current research programme "Reading and Writing Strategies of Disabled Groups" and comprised 5 production tasks: a picture story task ("Frog where are you?") based on a wordless picture story booklet by the American artist Mercer Mayer [7], a narrative task with a pre-set topic ("I was never so afraid"); a route direction task; a job application task; and a "letter to the editor" type of argumentative discourse task. All writing activities were computer-logged by means of the computer tool "ScriptLog" [8].

3.2. Scriptlog

Scriptlog is a simple editor that keeps a record of all events on the keyboard (i.e., the pressing of alphabetical and numerical keys, cursor keys, the delete key, space bar etc, and mouse clicks), the screen position of these events and their temporal distribution. From a Scriptlog record, you can then derive not only the finally edited text from a writing session, but also the ”linear” text with its temporal patterning, pauses and editing operations and log file with detailed information about each keystroke. Examples 1 and 2 below shows a fragment of a finally edited text (with english translations in italics) and a correspondent linear fragment.

(1) "ALDRIG HAR JAG VARIT SÅ RÄDD"
I WAS NEVER SO AFRAID
Aldrig har jag varit så rädd som när jag senaste gången rökte i min fars bil.
I was never so afraid as the last time I smoked in my fathers car.

(2) <START><95.92>ALDRIG HAR GA<2.93>G <DELETE4>JAG VARIT SÅ RÄDD<4.82>2<DELETE>2<DELETE> <4.25>"<4.98><MOUSE(29,0)><2.18>" <5.73> <MOUSE(1,30)><3.20><CR><CR> <154.22> Aldrig har jag varit så rädd<15.07> <DELETE20> r jag varit så rädd som när jag rökte i min fars bil sist<5.20> <DELETE4> <6.53><DELETE21> senaste gången rökte i min fars bil<2.47>.<CR>

Example 1 shows the heading and the first sentence of a narrative written by a control subject. Example 2 shows the linear text as follows: First the start button is pressed. After that the writer waits 95,92 seconds before starting to write - perhaps thinking about how to start, or planning the plot. Then she writes ALDRIG HAR GA and pauses for 2.93 seconds. She writes another G and a space, and then notices that the last word was wrong. She deletes four tokens (the space and GAG) and writes JAG (I) instead. And so on. <MOUSE(29,0)> means that the writer moves 29 tokens backwards in the text and <CR> means Carriage Return (new line).

3.3. The Corpus

The corpus consists of 149 edited texts containing altogether 39527 words. Table 1 shows an overview of how the words are distributed over the three groups. However, as was mentioned above the finally edited texts don't show everything that has been written but only those parts which the writers have chosen to keep. Therefore the total number of keystrokes it has taken to produce these texts and the final numbers of characters left in the texts are also shown. The table further presents the number of pauses and editings which have been used for the analyses in this paper. The frog story is excluded from many analyses and the numbers in italics show the relevant numbers for the other four texts taken together.

Table 1: Overview of the corpus

Group	keystr. lin	keystr. fin	words fin	pauses	ed
Norm	115337	97314	16911	7023	
excl frog	*66734*		*9290*	*2576*	*1167*
Dys	77963	61292	11854	12461	
excl frog	*46695*		*6720*	*4815*	*1212*
Deaf	73894	57108	10762	4311	
excl frog	*33445*		*4749*	*1095*	*690*
Sum	267194	215714	39527	23795	
excl frog	*146874*		*20 759*	*8486*	*3069*

4. Analyses

Three types of disfluencies were found. As expected unfilled pauses are the most common but also repairs are also frequent. Only quantitative analyses of these, which could be made automatically are presented. The pros and cons of this are discussed in section 5. However, in the texts produced by the deaf subjects a phenomenon that resembles filled pauses also occur. The analyses of these three types of df:s are presented in the following sections.

4.1. Unfilled pauses

4.1.1. Pause definition

By a pause, is here meant a transition time between two keystrokes which is longer than what can be expected for merely finding the next key. To make a pause a writer has to interrupt his/her typing considerably longer than the normal transition time between two keystrokes. Therefore pause criteria should optimally be individually tailored according to each subjects' individual typing speed.

However, like in some earlier studies of writing (e.g. [9]) it was stipulated that all transition times longer than two seconds should count as pauses. This is more than twice the median transition time between two letters for our slowest subject (0.817 sec). The subjects in the three groups type with different speeds. The deaf are in general the fastest and the dyslexic group the slowest.The advantage of the two second criterion is that it is quite safe to assume that transitions counted as pauses really are pauses. The disadvantage is that some pauses of the faster writers may not be included.

Pauses are categorized according to which type of *microcontext* they occur in.Each transition between two keystrokes is a possible pause location and a pause could for example occur in the transition between two lower-case letters as just mentioned, or between a space and a letter etc. Table 2 describes the notation used to describe microcontexts. According to this 'a*a' describes a pause that occurs between two letters, and a_*a describes a pause that occurs between a space, which follows a letter, and a letter etc.

Table 2: Notation used to describe microcontexts

character	description
a	a letter
_	a space
*	a pause
D	a deletion
. and ,	major and minor delimiters

4.1.2. Overall pause frequencies (pauses > 2 sec)

Table 3 shows the frequency of pauses in subcorpora of the three groups, with and without the frog story.

Table 3: Pause frequencies

Group	Pause/word		Pause/keystroke	
	Incl frog	Excl frog	Incl frog	Excl frog
Norm	0.359	0.305	0.051	0.041
Dys	0.895	0.753	0.135	0.108
Deaf	0.378	0.320	0.050	0.040

First, we observe that for all groups pauses in writing are more frequent than all disfluencies taken together in speaking. See for example [10] for an account of df:s in Swedish dialogues. Second, the pause frequencies were compared both for group and for text types. There was a significant effect for both ($p< 0.0001$). All groups make more pauses when writing the frog story than when writing the other texts. The explanation for this is probably that the subjects look at the pictures a lot. The frog story is therefore excluded from the rest of the pause analyses. Further, the dyslexic subjects make many more pauses than the other two groups . This was expected, both due to their writing problems and to their lower writing speed.

4.1.3. Pause lengths

First, the overall pause lengths were compared for group and text type. There was no effect for group alone and only a weak effect for text type. However, there was a significant effect for group and text type together (p=0.0075). The deaf subjects made much longer pauses in the "letter to the editor" texts than in the other text types. It is difficult to find a good explanation for this, but perhaps they had little experience of writing argumentative discourse.

Second, the lengths of pauses were compared for different microcontexts. Table 4 shows the results of this analysis.

Table 4: Mean Pause length (pauses $>$ 2 sec)

Context	Norm		Deaf		Dys	
	n	mean len	n	mean len	n	mean len
a_*a	506	5.2	213	4.8	1152	4.7
a*a	101	3.3	93	5.2	736	4.1
a*_a	252	5.6	187	4.2	579	5.2
._*a	112	6.9	55	5.1	184	6.4
a*.	103	7.6	90	3.9	180	7.5
.*_a	51	6.7	34	5.3	57	11.6
,_*a	35	6.7	8	7.0	18	5.4
a*,	72	4.3	26	4.0	52	10.7
,*_a	4	7.1	3	3.7	5	4.6
a*D	259	7.5	65	4.4	500	8.0
D*D	0	–	39	7.9	214	8.1
D*a	98	6.1	58	3.7	500	6.0

The table is divided into four parts. The first part deals with microcontexts that are likely to occur in the beginning of, within and the end of a word. The second deals with microcontexts which are likely to occur in the beginning and end of a sentence, and the third deals with microcontexts associated with units which are delimited by a comma. The fourth part deals with microcontexts associated with deletions and will be discussed in section 4.3. The table shows how many pauses of each context occur in the corpus and their mean length. The table shows that for the control group and the dyslexic group the mean pause lengths tend to be longer for the sentence related pauses than for the word related pauses. However, surprisingly this effect is significant only for the dyslexic group (p=0.0001). There are several possible explanations for this. First, it may be due to the pause criterion. If shorter pauses were included for the faster typists the means may look different. Notice that the differences in amounts of pauses are much higher for the word related pauses than for the sentence related pauses. Another explanation may be that the pause categories need to be more fine-grained. Perhaps we would see another pattern if we also marked phrase boundaries in the texts. The deaf group show an even more even distribution than the other two groups. It is impossible to tell if this is due to any special characteristic of the group or if the same explanation holds for the deaf as for the normal writers. Nothing specific was found for the comma related contexts.

4.1.4. Pause frequencies in certain microcontexts

Let us now turn to the frequencies in the same microcontexts. Table 5 shows for each context the percentage of pauses in each microcontext and how many of the subjects in each group had included this type of context in their texts at all. For example only eight normal subjects appear to have written `a*,` and 36.6% of these microcontexts cooccurs with a pause $>$ 2 seconds.

Table 5: Proportions of pauses $>$ 2 sec in certain contexts

Context	Norm		Deaf		Dys	
	mean perc	no subj	mean perc	no subj	mean perc	no subj
a_*a	7.1	10	4.8	9	18.9	11
a*a	0,3	10	0.5	9	3.4	11
a*_a	2.8	10	4.3	9	9.2	11
._*a	31.3	10	21.4	9	55.1	11
a*.	22.3	10	23.8	9	52.7	11
.*_a	12.4	10	17.1	9	19.6	11
,_*a	11.1	10	32.1	8	25.7	9
a*,	36.6	8	27.0	8	84.7	9
,*_a	2.5	8	14.7	8	3.3	9
a*D	26.4	10	13.8	9	50.8	11
D*D	0.0	10	1.4	9	5.8	11
D*a	11.3	10	10.0	9	54.1	11

Although pause lengths didn't increase with unit size, the sentence related microcontexts appear to cooccur with pauses much more often than the word related pauses. This result appears to hold for all groups, and perhaps it could partly explain the results of the length analysis. Notice the very high percentage of pauses between a letter and a comma for the dyslexic group. One possible explanation for this may be that they often use comma instead of full stop.

4.2. Filled pauses?

As has been mentioned earlier, filled pauses are not expected to occur in written monologues. However, in the texts of the deaf group, something which resembles filled pauses occur. They use series of dots ranging from two to five full stops in a row. 392 occurrences of the microcontext `.*.` were found in their texts. The corresponding numbers for the normal writers were 12 and 0. These series appear as units. There are no pauses within them. Eight of the nine deaf subjects use these patterns.

A possible explanation for this result could be that it is a text telephone convention and that they are influenced by that, since their main use of writing is in text telephone conversations.

4.3. Repairs

There are different strategies to make repairs in writing. Obviously they don't begin with an explicit editing phrase like they sometimes do in speaking. The two main repair strategies is to start deleting from the current character insertion point and to move to another point in the text and make a deletion and/or an insertion. More than 90% of all repairs in this corpus involved deletions. By measuring the editing range (how much is deleted) and the editing distance (how far the writer moves to get to the repair point) we can get an idea of how a writer makes repairs.

Table 6 shows the mean editing frequency, the mean editing range, the proportion of editings with a range of only one letter and the mean editing distance for the three groups.

Table 6: Deletions: frequency and range

Group	Ed/key	Ed/word	Range	1-let	Dist
Norm	0.014	0.136	3.262	55%	4.39
Deaf	0.017	0.165	3.408	44%	0.13
Dys	0.021	0.188	3.101	54%	3.92

Like for pause frequencies we observe that the frequencies of repairs are higher in writing than in speaking. All of these, except the editing distance were compared for group and for text type. There was no difference between the groups for any of these variables , but an effect for text type was found for editing frequency ($p=0.0128$). For some reason, all groups made more repairs in the route descriptions. It is worth noticing that about 50% of the repairs are changes of only one letter. These are due to either spelling problems or typos. Concerning the editing distance, the variance between the subjects is too large to make any reliable statistical analyses. Many subjects have total editing distance which is zero. Among the ten normal writers six subjects have an editing distance > 0 in 16 texts altogether. For the dyslexic subjects the correspondent results are nine subjects in 25 texts and for the deaf only two subjects in four texts.

Let us finally look at how pauses and repairs interact (see table 4). While there were no great differences between the groups in how long pauses they made in word, sentence and comma related microcontexts, we find that the deaf subjects appear to make shorter pauses than the other two groups before and after repairs. These results agree with their short editing distance. However, within repairs both deaf and dyslexics make quite long pauses while the normal writers don't make any pauses longer than two second at all. Perhaps this resuts could be explained by their writing problems for the dyslexics and L2 situation for the deaf.

5. Summary and conclusion

To sum up unfilled pauses apppear to be the most frequent df in writing. Further, pauses and repairs appear to be more frequent in writing than in speaking. Considering the different production conditions between spoken and written language this was an expected result. As in spoken language production sentence boundaries cooccur with pauses more often than word boundaries. These results holds for all groups and could therefore be considered as quite robust. Perhaps it could be suggested that while the inter-individual functions of df:s are different in speaking and writing, the intra-individual are quite similar.

Concerning the specific characteristics of the the two groups with disorders, the dyslexic subjects are more disfluent than the normal writers. This was expected, considering their spelling problems. They have a higher pause frequency than the other two groups in general, and specifically a high proportion of pauses within words. Surprisingly they were the only group who showed a significant difference between the pause-lengths associated with sentences and words. It is suggested that this result is an artefact of the high common pause criterion and the fact that no intermediate constituents were analysed. A manual tagging for phrases and clauses within the sentences are suggested as further work. A methodology for setting individual pause criteria, related to each writer's typing speed is also needed.

The deaf subjects on the other hand are not specifically disfluent and their writing process is quite linear. They are fast typists and don't often move far in the text to make a repair. Interestingly, a specific characteristic of their text is that a phenomenon, which resembles filled pauses, sometimes occur in their production. These results could perhaps be explained by their frequent use of text telephone, where writing is used under on-line production conditions, and it is suggested that the on-line/off-line distinction is at least as important as the modality distinction for the language production process. A way of investigating this issue further would be to make on-line recordings of text telephone conversations.

6. References

[1] Gregg, L.W. and Steinberg, E.R., Eds, Cognitive processes in writing, Hillsdale: Lawrence Earlbaum Associates Inc. 1980.

[2] Hayes, J.R. and Flower, L.S. "Identifying the organisation of the writing process" , In: Cognitive processes in writing, Gregg, L.W. and Steinberg, E.R., Eds., Hillsdale, NJ: Lawrence Earlbaum Associates, 1980, pp 3-30.

[3] Lickley, R. and Bard, E. "On not recognizing disfluencies in dialogue", Proceedings International Conference on Spoken Language Processing, Philadelphia, PA, Vol 3, 1996, pp 1876-1879,

[4] Clark, H. "Speaking in time", Proceedings of the ESCA workshop on dialogue and prosody 1999, pp. 1–6.

[5] Allwood, J., Nivre, J. and Ahlsén, E. "Speech Management — On the Non-Written Life of Speech", Nordic Journal of Linguistics, Vol 13, 1990, pp 2–48.

[6] Swerts. M. Filled pauses as markers of discourse structure", Journal of Pragmatics,Vol 30, 1998, pp 485–496

[7] Mayer, M., Frog, where are you?, New York: Dial Press, 1969

[8] Strömqvist, S. and Malmsten, L. "ScriptLog Pro User's manual", Gteborg University, Dept of Linguistics, 1998

[9] Spelman Miller, K. "Academic writers on-line: investigating pausing in the production of text", Language Teaching Research, Vol 4, No 2, 2000 pp. 123–148.

[10] Eklund, R., "A comparative analysis of disfluencies in four Swedish travel dialogue corpora", Proceedings of Disfluency in Spontaneous Speech Workshop, Berkely, Californa, 1999, pp 3–6.