www.sltinfo.com

ABSTRACT: The type-token ratio (TTR) is a measure of vocabulary variation within a written text or a person's speech. The type-token ratios of two real world examples are calculated and interpreted. The type-token ratio is shown to be a helpful measure of lexical variety within a text. It can be used to monitor changes in children and adults with vocabulary difficulties.

Type-Token Ratio of Written Language

Take a look at Text 1 below, which is an extract from something I wrote a while ago.

> **TEXT 1: Written Language**
>
> But what are thoughts? Well, we all have them. They are variously described as ideas, notions, concepts, impressions, perceptions, views, beliefs, opinions, values, and so on. At times they are brief, coming and going in an instant. On other occasions they seem to endure and we can mull them over again and again in the act we call *thinking*. We can put them aside, fall asleep, and then return to them later. We refer to them as things we can handle. However, this is just a metaphor.

If we count the number of words we get a total of 87. The number of words in a text is often referred to as the number of *tokens*. However, several of these tokens are repeated. For example, the token *again* occurs two times, the token *are* occurs three times, and the token *and* occurs five times. The following table shows all the tokens in Text 1, together with their frequency of occurrence.

rank	word	freq	rank	word	freq	rank	word	freq	rank	word	freq
1	we	6	17	asleep	1	33	impressions	1	49	seem	1
2	and	5	18	at	1	34	instant	1	50	so	1
3	them	5	19	beliefs	1	35	is	1	51	the	1
4	are	3	20	brief	1	36	just	1	52	then	1
5	can	3	21	but	1	37	later	1	53	things	1
6	they	3	22	call	1	38	metaphor	1	54	thinking	1
7	to	3	23	coming	1	39	mull	1	55	this	1
8	again	2	24	concepts	1	40	notions	1	56	thoughts	1
9	as	2	25	described	1	41	occasions	1	57	times	1
10	in	2	26	endure	1	42	opinions	1	58	values	1
11	on	2	27	fall	1	43	other	1	59	variously	1
12	a	1	28	going	1	44	over	1	60	views	1
13	act	1	29	handle	1	45	perceptions	1	61	well	1
14	all	1	30	have	1	46	put	1	62	what	1
15	an	1	31	however	1	47	refer	1			
16	aside	1	32	ideas	1	48	return	1		**TOTAL**	87

We see, then, that of the total of 87 tokens in this text there are 62 so-called *types*. The relationship between the number of types and the number of tokens is known as the *type-token ratio (TTR)*. For Text 1 above we can now calculate this as follows:

> **type-token ratio** = (number of types/number of tokens) * 100
>
> = (62/87) * 100 = **71.3%**

The more types there are in comparison to the number of tokens, then the more varied is the vocabulary, i.e. it there is greater *lexical variety*.

Type-Token Ratio of Speech

Now take a look at Text 2. This is an extract from a transcribed conversation between two people, P and A.

TEXT 2: Speech

```
01    P:    so: (.) er: (..) as you were saying about er:: (.)
02          where are you living now Andrew
03    A:    Skipton Lodge
04    P:    Skipton Lodge?
05    A:    mm (...) Skipton Lodge
06    P:    yeah (.) do you like it
07    A:    yeah I do
08    P:    yeah
09    A:    I've settled in
10    P:    you have (...) good (.) w w what are the things you
11          like about it
12    A:    go out in the tow:n
13    P:    you go out in the town (...)
14    A:    yeah
15          (2.1)
16          with er: Tommy and Martin (.) and er:: (.) Noel
17    P:    and?
18    A:    NOEL
19    P:    oh yes (.) oh he lives there does he?
20    A:    yeah he live(s) in the flats
21    P:    yeah (.) oh they have flats there do they
22    A:    mm
23          (3.3)
24          and er::
25          (2.3)
26          and I went to see (..) (Elaine)
```

As before, I have set out the tokens and their frequency of occurrence in tabular format in the Table below (I have ignored pauses such as (…), (2.1), the repetition of initial /w/ in line 10, and the *inserts er, oh* and *mm*).

rank	word	freq	rank	word	freq	rank	word	freq	rank	word	freq
1	yeah	6	13	flats	2	25	as	1	37	to	1
2	you	6	14	go	2	26	does	1	38	Tommy	1
3	and	5	15	have	2	27	Elaine	1	39	've	1
4	in	4	16	it	2	28	good	1	40	went	1
5	the	4	17	like	2	29	living	1	41	were	1
6	do	3	18	lives	2	30	Martin	1	42	what	1
7	he	3	19	Noel	2	31	now	1	43	where	1
8	I	3	20	out	2	32	saying	1	44	with	1
9	Lodge	3	21	there	2	33	see	1	45	yes	1
10	Skipton	3	22	they	2	34	settled	1			
11	about	2	23	town	2	35	so	1		**TOTAL**	**88**
12	are	2	24	Andrew	1	36	things	1			

We can now calculate the type-token ratio as before:

type-token ratio = (number of types/number of tokens) * 100
= (45/88) * 100 = **51.1%**

Interpretation

You will see that the number of tokens in each of the texts is almost the same (87 in Text 1 and 88 in Text 2). However, the type-token ratios are different: 71% for the written text (Text 1) and just 51% for the spoken text (Text 2). We can say, therefore, that the vocabulary is less varied in the spoken text than in the written text. Or, to put it another way, the written text shows greater lexical variety. A high TTR indicates a large amount of lexical variation and a low TTR indicates relatively little lexical variation. This finding, that the type-token ratio of speech is less than that of written language, is typical.

A major difference between speech and written language is that speech in conversation is produced in real time. There is limited time to think about, and plan, what one wishes to say. Consequently, speakers tend to select words from a relatively restricted vocabulary. In contrast, an author of a written text has much more time to plan and select just the right vocabulary items that best communicate his or her meaning.

As with *lexical density*[1], the type-token ratio can also be used to monitor changes in the use of vocabulary items in children with under-developed vocabulary and/or word finding difficulties and, for example, in adults who have suffered a stroke and who consequently exhibit word retrieval difficulties and naming difficulties.

---

[1] Lexical density is a measure of how much information is contained within a text.
See http://www.sltinfo.com/lexical-density.html