LN2000 - Language and the Computer

Todor Arnaudov

06xxxxx

My corpus about HDD-RAM-CPU

1. Is it really a corpus?

My collection of texts is a corpus, because it fulfils all the requirements for being one:

I. The articles are naturally occurring texts, taken from two web sites: Tom's Hardware and TechReport - resources for deep reviews of computer hardware.

II. My collection has finite size, and, of course, is in machine readable form.

III. I can share the original texts I've collected, on demand.

I selected three domains and collected approximately equal amount of texts from each one:

1. Storage - reviews and analysis of hard drives and DVD drives.
2. Memory - Random Access Memory units.
3. Processors – mostly reviews and comparisons of recent Intel and AMD's CPUs.

The articles about Storage give information about new developments in hard drive industry, such like the review of the newest Hitachi Terabyte PC Hard Drive and its amazing performance, compared to other recent devices. There are also reviews and performance comparisons of recent hard drives at different size-factors – 3.5”, 2.5” and 1.8” - and a text which explains how the modern hard-drives work.

The part, devoted to memory consists of reviews of DDR modules and a text about the history of RAM modules for PC from '80ies until beginning of 2000s. “The memory war” between late 90ies and early 2000s is mentioned, when several standards caused a lot of compatibility issues.

One of the most famous "battles" in the “memory war”, mentioned in the corpus, is the commercial fault of Intel, when they choose expensive and badly supported Rambus RDRAM modules to be the only compatible with the first motherboards, created for Pentium 4 CPUs.

Processors are covered by reviews and comparisons of Core 2 Duo, AMD Athlon 64 X2, and AMD Turion 64 X2 mobile processors, and with a story about different socket standards of AMD processors.

2. Building the corpus

First, I selected a genre and sources – articles about computer hardware from the Internet. I wanted to create a corpus in a specialized area, in order to display the frequent use of terms. I didn't ask for permission - similar analysis could be done and probably is actually done by crawlers of search-engines. Also, the use of these texts should be considered fair - purely educational; I don't think anybody from the authors will be angry to me for that use.

Then, I collected texts in the desired topics with copy-pasting in several text files on my computer - for easier access and pre-processing.

I considered the texts didn't require spell-checking, since they were quoted from famous sources, where such kind of mistakes are supposed to be very rare. I did a bit of cleaning, though, because while I've been pasting texts from the pages, some lines came two times – that happened with captions under pictures and diagrams, which came together with the same text of the “alternative” tag behind the image. Also, the articles were divided in a lot of sub-pages with the title repeated over and over again – I've been pasting the titles just once, and I've omitted the captions “Continued” and the links “Previous, Next”, before and after the body of the texts.

While I've been storing the texts in files on my computer, I've also been uploading them to the on-line corpus-service on-the-fly, in order to evaluate the corpus in development. For example, in the beginning after I was filling the corpus with articles only from the domain of hard-drives, the word “drive” appeared as high as 7th place; after adding variety of texts, it went down to 16th place

and was left behind several general auxiliary words and two term nouns: “memory” and “GB”.

3. Evaluation

I built the corpus with the goal to display that when our texts are in specialized genre, we'll see particular terms to show up high in the frequency list.

Rank	Frequency	Word
1	3526	the
2	1553	to
3	1424	a
4	1420	of
5	1372	and
6	877	is
7	777	s
8	722	in
9	684	for
10	579	that
11	543	The
12	530	**memory**
13	507	with
14	492	**GB**
15	475	at
16	465	**drive**
17	444	are
18	414	as
19	410	on
20	394	**drives**
21	387	you
22	367	**MB**
23	356	it
24	333	**DDR**
25	330	we
26	315	be
27	307	**MHz**
28	295	**performance**

We may clearly see several terms ranked very high: **GB, MB, DDR, MHz; drive, drives**. The appearance of the noun **"performance"** is explained by the nature of the articles, since most of them are performance evaluations.

We find "the missing" CPU domain few lines below, but ranked very high, also; as expected:

Top occurrences from CPU domain

35	**230**	**Intel**
46	**205**	**Core**
59	**168**	**Athlon**
60	**165**	**clock**
64	158	time
65	**157**	**CPU**
66	**156**	**GHz**
67	147	PC
72	**132**	**processors**

CPU-domain emerge clearly by looking at the 2-grams, as *Athlon 64* and *Core 2* appear at 4th and 5th place, respectively. The multiword *expression "hard drive"* is now 8th.

At 3-grams, processors' domain occupies **all top 10 (!) places**, which shows that multiword expressions are used very frequently, because CPUs are often quoted by their full names, and because the competitors in PC CPUs are very few, compared, for example, to the variety of competing hard drives on the market. The first "regular" 3-gram is at 11th place (in order to), but some of the interesting "regular" 3-grams are at 30th and below: "a look at", "As you can [see]",- they display that this corpus consists of reviews. "Hard drive" is splitted in several locations – it's just two-word long, that's why as a term it fells down in the list.

24	17	the hard drive
65	11	hard drive to
84	10	a hard drive

Relatively Equal distribution of words from all domains appears as late as 5-grams, when in top-ten we see terms from all:

1	**18**	**the Athlon 64 X 2**	**CPU**
2	**15**	**DDR 2 SDRAM at 800**	**RAM**
3	**14**	**Native Command Queuing (NCQ**	**HDD**
4	12	Athlon 64 X 2 5000	CPU
5	12	2 GB (2 DIMMs	RAM
6	11	s 7,200 RPM 16 MB	HDD
7	11	Core 2 Extreme X 6800	CPU
8	11	Athlon 64 X 2 4800	CPU
9	10	Barracuda 7200.10 Barracuda 7200.10 Barracuda	HDD (probably from a table)
10	10	Athlon 64 X 2 6000	CPU

Comparisons between my corpus and BNC

The first 11 items in the frequency list seem to represent a distribution, which is very similar to BNC. However, on 12th place we see "memory" , which is out of top 1000 in BNC. Same goes for "drive" and, of course, all specific memory, HDD, CPU and frequency terms. In general, due to the richness of terms, the corpus consists of immensely more nouns in the top positions.

We could note an interesting similarity between distributions of "time". In my corpus it's slightly more frequent than in BNC (64th vs. 69th), but the uses in BNC are usually different and, generally, much more diverse.

provide help in many different ways to ensure that people don't spend **time** in hospital unnecessarily.
How much **time** to I need to give?
It was the first **time** our national and international network
The author has had **time** to consider and reflect, so that descriptions
Only **time** and scholarship eventually sort out the various relationships

In my corpus, "time" is very often preceded by an adjective, as it's a part of terms:

An **access time** of 13.8 ms is an
without having a **wait time**
average **seek time** in its published specifications
DVD creation time was quick, but the
the chip manufacturer in **cycle time** (measured in nanoseconds

Not surprising, but notable is that the pronominals **"she"** and **"he"** are almost absent in the corpus - just 7 occurrences out of 93000 words. That's far away from top 1500, whereas in BNC they are ranked between 20th-50th place. On the other hand, **"I", "you", "we"** and "**they**" have similar distributions (55/12, **21/18**, 25/46, 89/37). The above is due to the genre of the corpus - while individual "third persons" are rarely referred in texts reviewing computer hardware, it's not unusual the author to refer to himself - he gives his opinions and decisions he has made, - and to express statements about groups of reviewed items, readers, users, companies etc.

Linked to the above is the rareness of the noun "people" - 1159th with just 9 uses, opposite to 91th in BNC, and the role of the word as the second noun there. In the texts from our corpus the authors address their texts directly to "you, the reader" (“you” is at 21th) and rarely mention "the people" or talk about “people” in general, because the corpus is about machines.

We'll finish our analysis here, even though there are a lot of other interesting things we could mention...

Todor Arndaudov, University of Wolverhampton

30/4/2007