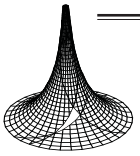

3/2 (2005), 203–219
tmcs@math.klte.hu
http://tmcs.math.klte.hu

Teaching
Mathematics and
Computer Science

Research on IT language use at a company

Attila Kálmán Juhász

Abstract. The aim of the research of the IT language, used in the written documents of a company, is to contribute to the creation of a (mono- or bilingual) dictionary or encyclopaedia available for the public on the Internet, serving, among others, as a reference tool for the unified, controlled and unambiguous use of IT terms for students at various educational levels. To this ongoing work, the participation and cooperation of a panel of experts of different competences, linguists as well as IT experts, is indispensable.

The methods of corpus linguistics were used to carry out the research. The IT terms were separated from the texts and then a concordance software was used to see the environment of the IT words and phrases in which they occur. So their morphological analysis became possible.

The results of the research showed that a great number of Hungarian morphological language use problems stem from the way the IT terms are used in the documents. This paper lists, groups, analyses these phenomena.

The conclusions of the author are: (1) If such an Internet dictionary is used generally and consulted when e.g. somebody wants to write a composition or essay, translate an article, write a newspaper article, a scientific publication or a textbook to be taught at schools of different types and levels, etc. most of the communication noises could be filtered out. (2) At the same time it could promote the use of adequate (both in linguistic and technical meaning) Hungarian terms eliminating the "Hunglish" usage. (3) It could also contribute to the prevailing use of the relevant Hungarian terminology. Such a dictionary would be indispensable, not only in educational and industrial environments but in the electronic and traditional media as well. Last but not least, it could raise the level of different teaching materials (textbooks, e-materials, etc.) used in public and higher education.

Key words and phrases: IT terms, corpus linguistics, KWIC concordance, Internet website, panel of experts.

ZDM Subject Classification: P80.

1. Introduction

1.1

The central problem of teaching informatics is the complexity of IT language.

We have to face a rapidly developing area. For instance, in the world of PCs, a model a few years old is hardly usable, and the "novelties" that came out a decade or so ago are practically out of date. It is naturally reflected in the language of information technology. The extreme popularity of computers – which being so widely spread, and which was unimaginable earlier, also has had a great impact on IT language. Nowadays computers reach all kinds of consumers, regardless of age, sex, qualification, position, occupation, etc. Due to this fact, users with different levels of knowledge have to meet an enormous amount of minor or major problems. In this profession asking for advice seems to be almost the only way to sort out a problem. Preferably it must be done face to face, but if it cannot be done in person you can correspond with a paper, magazine, or journal specialized in computer science, search the Net, or turn to an Internet forum for help.

Consequently, IT language, as the base for communication among computer users, has a very special significance compared to the other branches of science. The constant flow of communication results in an amazing abundance of IT language, with the unchecked as well as the uncheckable "flourishing" of IT terminology.

Higher education must pay special attention to the above-mentioned features of these phenomena and adapt to the situation. Practically, there is only one effective solution to this problem: the creation and continuous development, update and upgrading of a mono- or bilingual (even multilingual) IT dictionary or encyclopaedia on the Internet that is not only up-to-date, but monitors and tracks the oral as well as written language variations, and which is available and accessible to everyone who wants to find their way in the "jungle of informatics".

In the paper below I provide a detailed analysis of the special language use and communication (using the methods of corpus linguistics) between users and experts within a major industrial company. It is the opinion of many, that the experiences and results drawn from this research can form a firm base for such an IT dictionary.

1.2

The *aim* of the corpus-based research was to explore the characteristic features of IT language use at a company (hereinafter the Company) in Hungary. The scope of the activities of the company in question comprises the provision of all kinds of IT services, mainly to it's subsidiaries in the heavy engineering industrial sector, which the Company itself belongs to. The research was restricted to the analysis of its written documents of various kinds, but in this article I intend to deal only with their Help Desk documents.

As mentioned before, in our rapidly changing world where the pace of technical innovation has become very fast – seen especially in the very swift growth of the IT industry – new and new terminology appears day by day. It seems to be hopeless task keeping pace with all this development unless we have access to a controlled and up-to-date source of information (e.g. to a dictionary of IT terms – functioning as a kind of knowledge base). Without it, as the overwhelming majority of the new terms are in English, one is first informed about them probably in English through different, mainly electronic, – the Internet, television, radio – channels. However, the effect and influence of conventional media, the newspapers, magazines, books, etc., must also be taken into consideration.

Due to this phenomenon, the terms are rarely translated into the local languages. Instead, they start being used in different languages in their 'native' manner, causing lots of problems in many respects to the given languages.

1.3

The use of the new vocabulary is problematic, with regards to *semantics*, as the meaning of a given and new term is not clear. Even if the people who start to use a new word or phrase are on one hand IT experts – or at least have a solid knowledge of the area – and on the other hand their knowledge of English is sufficient to understand ESP language as well, confusion is still possible. We should not forget about the masses of people, including the majority of students starting their studies, who are lacking these skills. For them the use of the original terms can cause confusion, not to mention the problems caused through the use of acronyms and abbreviations.

It must be emphasized that the creation and use of proper metaphoric terms is essential. According to the modern concepts on metaphors, they play an essential role in the appearance of new concepts, as well as in the process of understanding the relationships among them (conceptualization). This is due to the fact that the

metaphors generally connect the new concepts (from the so-called target domain) to the previously known ideas (that is, the source domain) [Boda–Porkoláb, 2000]. Using foreign (e.g. English) terms without their native equivalents, therefore totally ignoring their rich associative relationships, the above-mentioned functions of metaphors cannot be fulfilled.

1.4

The use of the English words causes *morphological* problems as well. In an agglutinative language like Hungarian, it is awkward to see English stems combined with Hungarian prefixes and/or suffixes which often leads to the violation of the orthographical rules as well.

1.5

Consequently, the rules of Hungarian *syntactics* are also affected.

All these contribute to the corruption of the mother tongue. No civilized nation can give up its demand for cultivating any science in its own scientific language. There are countries that set a good example of how to protect their mother tongue: Iceland and Turkey, among others. In these countries a carefully selected board of scientists, consisting of IT experts as well as linguists, create and define the new terms in their mother tongue as soon as they appear. The terms spread fast through all modern media, and from that point they are published in that form, and people start using them both in oral and written communication. In this, I can see the signs of a consequent and successful language planning. The same attitude could be applied in Hungary as well. A successful implementation of this idea could be the creation of a special website, where a controlled IT dictionary would be beneficial.

The time factor is important, and the new words have to be competitive, too. They must sound Hungarian, preferably they should not be longer than the English version and of course they must adequately reflect the technical meaning. Everybody could benefit from such a practice, as it could eliminate the chaos that now can be observed in relevant textbooks, teaching materials, software descriptions, and the formal language use in offices, at schools of all levels, on the Internet and in private discourse as well, just to mention a few areas.

2. The methods of the research

2.1

I found the methods of *corpus linguistics* – that form part of *Natural Language Processing (NLP)* – suitable for analysing the information technology language use at the Company as corpora can provide the basis of accurate, empirically justified linguistic observations.

2.2

Corpus linguistics has made enormous progress since it started and initially was used for describing spelling conventions, language acquisition and language pedagogy.

2.3

The corpus of the "Company" contains different sub-corpora, one of which is the "Help Desk Documents". It is not a finite size corpus. I chose the method of a *monitor corpus*, as I did not want a "snapshot" type of synchronic research, for the corpus is not big enough to provide me with all the representativeness of the IT language use (however, it is representative with respect to the Company). Therefore, by collecting the documents of the Company I could construct a reasonably large and broad sample.

2.4

The corpus is an annotated one. Annotation here simply means that all words of the corpus were parsed ("raw material", the complete text, more precisely series of texts of different types).

Processing was carried out in different steps. First, the raw material was lemmatized and part-of-speech parsed. For morphological parsing a *unification-based word grammar parser* was used that was developed to correct the deficiency of older version of parsers. They could tokenize a word into a sequence of tagged morphemes but they could not determine the part of speech of a word or its inflectional categories (in a unification-based word grammar parser a third analytical component – the word grammar – was added to the rules component and the lexical component, or lexicon).

All this was carried out by using the expertise and the hardware and software capacities of Morphologic. There were consultations with them with respect to the method and the results. Thanks to the multi-step process, and the acknowledged and proved quality of the software used by them, the reliability of the part-of-speech parsing was high; the efficiency of parsing and selecting IT terms also produced excellent results (100%). So eventually I had to pick the word I intended to investigate, that is the relevant IT terms.

2.5

It soon became evident that a *problem-oriented tagging* would best serve my research goal. To solve the problem – separation of IT elements from the "rest" – was not easy. A reliable method had to be found. I focused on (orientation) the important, from the view of the research IT, words and decided to label them (tagging). For this special software Morphologic) was applied.

To reach my goal, I had to separate irrelevant "general Hungarian" from the IT ESP language. With the help of Hungarian-English and English-Hungarian IT electronic dictionaries, and by applying special software (Morphologic), filtering out the IT vocabulary was done by running them against the whole corpus (see: acknowledgements). During the filtering procedure, the act of filtering was repeated using different IT dictionaries (min. 50 000 words each). Apart from that, all IT items were filtered out from the most comprehensive general language dictionary, and after its running against the corpus, what remained was practically a blank sheet, that meant that the corpus contained no other words but names of persons and companies.

2.6

After this, I used the Simple Concordance Program (SCP) and applying a *KWIC concordance* I was able to comfortably analyze the semantics, morphology and syntactics of the IT language use. This programme can be downloaded free from the Internet (`http://www.scp.com`). Its main function is to produce a concordance of words from texts. This is called KWIC, Keyword in Context. As I wanted to see and analyse IT words in context, to see their morphological and syntactic features, the use of this programme seemed – and proved – to be especially good at this. It can be used for educational purposes as well, for example one can instruct the programme to select IT ESP words from a text

to show their frequency of occurrence and the environment of their occurrence. With the help of SCP one can make different statistics etc.

3. The results of the research

3.1

The Help Desk documents appear in Excel format at the Company. Before processing the sub-corpus contained 11 536 lines and 126 354 words. After lemmatisation a part-of-speech parsing has been done and the corpus looked like this (short example taken randomly):

laptop[FN] {nem[FN], nem[HA]} indul[IGE]+period[PUNCT]
kontakthiba[FN]+period[PUNCT]

(FN means (N)oun; HA means (A)dverb; IGE means (V)erb)

There is a hesitation in parsing with the Hungarian word: *nem*. In Hungarian it can either be a noun, meaning *gender* or an adverb, meaning *no*. In this case it means the latter.

Original text (with its translation):

Laptop nem indul. (Laptop won't start.)
Kontakthiba. (Contact failure.)

After a thorough examination of the parsed text it became evident that the Hungarian words were used correctly in terms of morphology, syntactics and apart from a few mistakes there were no problems with spelling either (most of them are pure mistyping). Sentences – consisting of Hungarian words only – were grammatically formed and can be regarded to be part of the Hungarian language. It is natural, in the case of Hungarian native speakers, to have a full language competence. As we see their performance is also complete.

So the use of the parser proved to be beneficial in determining that there were no problems in the usage of Hungarian general language.

However, at first sight of the parsed text, the use of IT terms proved to be very problematic. It became evident and inevitable to select the ESP vocabulary and examine them morphologically. It was obvious from the first sight that the uses of English terms with Hungarian affixes do not cause syntactic problems, but their morphology shows very unusual and mixed variations contradicting the spelling rules of the Hungarian language.

Originally there were 126 354 words and after filtering out the obviously general language vocabulary 2601 words were left. After the final filtering, 465 keywords out of 2601, a conclusion can be drawn that only a small proportion of the words are ESP words. It must be considered that the genre of the texts at the Help Desk, being less technical, probably differ considerably from the language of a manual or a software description.

3.2

The detailed analysis of the corpus appears in the results of the KWIC concordance program within the Simple Concordance Program (SCP).

The project statistics supports the previous statements which I made in the last paragraph of 3.1. The keywords represent 9.88% of the total word count and only 17.87% of the whole vocabulary. See:

Project statistics:
Analysis based on the keywords
Total vocabulary = *2601 types*
Project word count = 129328 tokens
Keywords selected = 465 types (17.87773933% of whole vocabulary)
Keywords word count = 12784 tokens (9.88494371% of total word count)
Types/tokens = 0.03637359
Types/sqrt(tokens) = 4.11262936
Yule's k = 217.10577971

And here is a sample of the keywords that exemplifies the word frequency in the text in decreasing number. See a selection of the 465 key words/2601:

1094 nyomtat, ... 102 server, ... 84 router, ... 76 user, ... 75 sw-t, ... 74 file, ... 72 gerinc, ... 68 overflow, ... 62 broadcast, ... 58 nyomtatni, 58 nyomtato, ... 52 lefagy, ... 47 file-t, ... 47 patch-et, ... 40 fdd, ... 38 be-patch-elt, ... 30 patch, ... 29 routert, ... 24 egere, ... 23 firmware-t, 23 resetel, ... 20 upgrade, ... 19 lemezt, 19 linken, 19 port, 19 uninstall, ... 18 nyomtatja, ... 16 driver-, 16 hub-ot, 15 driverek, 15 serve, ... 14 domain, ... 13 reset, ... 11 serverrel, ... 10 kinyomtatott, 10 ujratelepitese, ... 9 linket, ... 8 ripiter, 8 routerekre, 8 servere, 8 szervert, ... 7 szerverre, 7 user-t, 7 winchesterhiba, ... 6 alaplapi, 6 boot-ol, 6 hub-ban, 6 inact-act-olni, ... 4 processzor, 4 szoftver, ... 3 adatai, 3 aljzatok, 3 aljzaton, 3 felismeri, *3 forgalom*, 3 kapcsolni, 3

karaktereket, 3 kazetta, 3 kikapcsolta, 3 kimenet, 3 kinyomtatni, 3 kurzor, 3 proci, 3 sorokat, 3 zavarokat, ... 2 adatbevitel, 2 adatkapcsolati, 2 adatokat, 2 busz, 2 cd-rw, 2 csatlakozott, 2 excelbe, 2 feljelentkezni, 2 html, 2 kliensek, 2 lapokat, 2 leolvasni, 2 monitoron, 2 parancssoros, 2 printerport, 2 rendszerre, 2 szalagja, 2 szerverhez, 2 szoftverben, 2 szoftvere, 2 tintapatront, ... 1 csatlakoztat, 1 csatlakoztatni, 1 lefagyott, 1 megjavult, 1 rendszeren, 1 riaszt, *1 visszadugtuk*.

So, the most frequently used keyword (IT term) was *nyomtat* (prints) that occurred in the whole project *1094* times. At the other end of the word frequency list we can find *visszadugtuk* (we have plugged it again) which appears only once in the corpus.

As the project statistics show there were 2601 types of words out of which 465 belongs to the group of ESP words, the keywords. I must make a remark here, that the IT terms are considered in a broader sense of the word term, as general language words like *forgalom* (broadcast) have also been selected because they have a well defined special IT meaning as well (see: conceptualization in 1.3).

Table 1. Word frequency profile of 465 words (bottom 10)

Word Frequency	**Number of Words**	**Cumulative Vocabulary**	**Cumulative Word**	**Percentage Vocabulary**	**Percentage Word**
1	7	7	7	1.50538	0.05476
2	25	32	57	6.88172	0.44587
3	33	65	156	13.97849	1.22028
4	22	87	244	18.70968	1.90864
5	11	98	299	21.07527	2.33886
6	48	146	587	31.39785	4.59168
7	38	184	853	39.56989	6.6724
8	29	213	1085	45.80645	8.48717
9	20	233	1265	50.10753	9.89518
10	19	252	1455	54.19355	11.38141

From the above word frequency profile (see: **highlighted row**) we can see that there were two keywords, each of which appeared in the project 215 times, 430 occurrences of these two words means 98.70968% out of the total 465 types

Table 2. Word frequency profile of 465 words (top 10)

Word Frequency	Number of Words	Cumulative Vocabulary	Cumulative Word	Percentage Vocabulary	Percentage Word
161	1	455	8312	97.84946	65.01877
165	1	456	8477	98.06452	66.30945
180	1	457	8657	98.27957	67.71746
215	**2**	**459**	**9087**	**98.70968**	**71.08104**
236	1	460	9323	98.92473	72.9271
294	1	461	9617	99.13978	75.22685
323	1	462	9940	99.35484	77.75344
798	1	463	10738	99.56989	83.99562
952	1	464	11690	99.78495	91.44243
1094	1	465	12784	100.00	100.00

of keywords. These two words belong to the most frequently (place 7 in top 10) used keywords of the KWIC project.

3.3

In this section a selection of keywords will be presented grouped around the typical morphological phenomena. To sample the KWIC concordance format a few examples will be presented, otherwise only the keywords will be shown with remarks and interpretation of the characteristic features of their use.

(a) As it was stated above, no morphological or syntactic problem can be observed in case of the Hungarian words. There will be remarks and interpretation mainly with regards to the English word forms. The KWIC concordance was arranged in alphabetical order of the appearance of the keys. The number in front of the key indicates the number of occurrences in the corpus. The second number refers to the line it appears in the text. Additionally I provide the English translation of the Hungarian terms and words in brackets, after the term(s) in bold characters.

2 **adatokat** (data) 247
5 **alaplap** (motherboard) 295
15 **bekapcsolás** (switch on) 781

13 **egere** 4959 >gépének **egere** rossz (the mouse of his/her computer is out of work)<

egere means: its mouse. Egér (mouse) is a mirror translation of the English word. This Hungarian term is deeply rooted in the language and it is exclusively used in Hungarian. It serves as a good example for the possibility of successful use of Hungarian terms. Its use was implemented from the very early days and has appeared in that form in writing as well as in speech ever since.

10 **ujratelepitese** 43641 >háló "**ujratelepitese**" (reinstallation of the net)<

újratelepítése means: its re-install. (*Újra* means re and *telepítés* means installation.) In the Hungarian IT "jargon" the term (újra)installálás also occurs. Both versions are alternatively used and this is a proof of a mixed, unsettled language use. The battle for survival in the language has not been decided yet.

(b) Five types of the use of the terms can be distinguished among the selected items: S1 stands for structure one, S2 is for structure two, and so on.

S1: 10 **bepatch-elni** 994 >Hálózatot kellene **bepatch-elni** és kábelezni (Net should be patched and cabled).<

A general structure in many cases is the Hungarian prefix + English stem + hyphen + Hungarian suffix(es). According to Hungarian spelling rules the use of the hyphen is incorrect.

S2: 7 **bepatcheltük** 14172 >**bepatchelt**ük. (we have patched it)<

It is the same as the previous one without the use of a hyphen. Prefix *be* (possibly means *in*, but rather refers to a kind of perfect tense here) + Stem + Infix *el* (verb inflection form, to form a verb from a noun) + *t* (Infl. form to indicate the past tense in Hungarian through the lack of explicit perfect tenses) + *ük* (to indicate first person in plural) All these types are very strange and look and sound awkward. Generally speaking the core of the morphological problems stem from the fact that English is an analytical language, while Hungarian is an agglutinative language with a very rich and varied set of inflections, suffixes, infixes, prefixes etc. to express person, time, mood, even (in)transitiveness etc. and for example case of nouns.

S3: 6 **be-patch-elt** 13792 >**be-patch-elt**ük.<

An unjustifiable version with a hyphen after the prefix *be*. Otherwise it is the same as S2.

S4: 6 **boot-ol** 1428 >**boot-ol**<

An example for a stem + suffix with an incorrectly applied hyphen meaning: He/she/it is booting or boots or has been booting depending on the context. In Hungarian language there are only three tenses. In this context it has a present progressive meaning.

S5: 62 **broadcast** 1487 >a rengeteg (loads of) **broadcast**<

It is an example of a frequent situation when the English words can be placed in a Hungarian sentence with no need (?) to change them. Then they cause neither morphological nor syntactic problem. They just sound or look strange and many times not understandable to the people who hear them or read them. *Broadcast* can be translated to Hungarian as *(adat)forgalom*. The Hungarian term is frequently used and also occurs in the corpus. But literally, *broadcast* means *közvetítés*, like TV or radio broadcast, and the English equivalent for *(adat)forgalom* is traffic or flow of data (*adat*). So the process of conceptualization and the creation of the concept have been carried out differently. The use of the Hungarian term is widely spread; still this English version is sometimes used. My suspicion is that it is due to snobbery in most of the cases.

The following samples represent S1–S5 variations. The randomly selected examples are in alphabetical order and arranged the same way as the previous ones.

S4: 8 **domain-be** 4828 >**domain-be** való bejelentkezést (registration on a domain)<
S4: 9 **file-jai** 42124 >**“file-jai”** (its files) A gép (The computer)<
S4: 6 **hub-ban** 8617 >leakadt (got unhooked) **hub-ban** (in the hub)<
S4: 16 **hub-ot** 8625 >ütközés (collision). A két 24-es **hub-ot** (The two hubs of 24)<
S5: 32 **driver** 4907 >Nyomtató **driver** csere (replacement of the printer driver). A hiba időszakos (Failure is temporary)<
S2: 15 **driverek** 4926 >hálókártya **driverek** ujratelepítése (reinstallation of net card drivers),<
S2: 10 **driverrel** 4938 >nyomtató driverrel a nyomtató nem (with the printer driver the printer won't)<
S2: 7 **drivert** 17364 >nyomtató **drivert** (printer driver)<

S2: 8 **e-mailok** 1742 >címre az **e-mailok** nem érkeznek (the e-mails don't arrive at the address)<

The suffix *ok does* not conform with the rules of vowel harmony. The diphtongue "ei" is not pronounced in Hungarian, instead it is pronounced "é", and after this a suffix *ek* must be used. The ungrammatical suffixation probably derives from a bad pronunciation of the word *mail*. Note that the use of *villámposta* is also widely used, though in this corpus there is no evidence of it.

S5: 90 **hdd** 6314 >**HDD** csere, adatmentés (HDD replacement, data saving)<

It is an example of the English acronym *HDD* (hard disc drive) for *merevlemez*. It is short and well established, though as we can see from the next example, the Hungarian version also can be found in the corpus:

4 **merevlemez** 42968 >semmire nem reagál merevlemez (hard disk reacts to nothing)<

S5: 236 **hw** 505 >Nem **HW** hiba. (not hw failure)<

It is the usual abbreviation for *hardware*, which is often spelt *hardver* in "Hungarian". Though many suggestions were made for its use in Hungarian but none has been widely established in the language. As it is one of the old basic words, and has been used in this form from the very beginning, it probably will become such a loan word which is going to be felt less and less strange. Some define it as something you can kick into, while *software* is something for which one kicks into the hw.

S5: 4 **szoftver** 44264 >is automatikusan leáll (also stops automatically) **szoftver** telepítése, (installation of the software)<
S5: 2 **szoftvere** 49002 >kábelt visszadugtuk (cable was plugged in again). **Szoftvere** nálatok (Its software at you)<
S5: 6 **szoftverek** 49005 >következő **szoftverek** telepítését kéri (asks for the installation of the following software):<

And finally here are some more examples that will be presented from the corpus without any more remarks:

S5: 102 **server** 2752 >Citrix **server** újraindítás, (reboot of the Citrix server)<

S2: 12 **serverhez** 46402 >Prioris **serverhez** tartozó (belonging to Prioris server)<
S2: 36 **szerver** 10163 >**szerver** állt. Újraindítás után (server stopped. After reboot)<
S2: 8 **szervert** 48989 >**szervert** kellett újraindítani (server had to be rebooted).<
S4: 6 **inact-act-olni** 8931 >megpróbálta **Inact-Act-olni** (tried to Inact-Act)<
S2: 19 **linken** 42608 >osztoznak egy 10Mbps-os **linken** (they share a 10Mbps link)<
S5: 68 **overflow** 45236 >folyamatosan **overflow** volt (there was a constant overflow).<

Before closing the series of examples and coming to the conclusions, let me draw attention to a special phenomenon, the so called "nickname-ization" of a – usually – longer word. *Processor* is spelt with double *sz* (the pronunciation of *sz* is the same as that of the *s* in English), that is, according to the rules sz+sz must be written ssz. The loan words of Latin origin do not sound strange in our language because of cultural and historical reasons. (See: *professor* is also spelt *professzor*.) But *processzor* is called *proci* which sounds friendly and is much shorter than the full word *processzor*. The same applies to the short form of *winchester*, i.e. *vinyó* or *vincsi*. (See the next example.) Letter *v* stands in the place of *w* as no *w* is pronounced in Hungarian just the vowel *v*. However, in writing, we always preserve *w*-s. But *vinyó* or *vincsi* is under the influence of the pronunciation and probably is felt to be a Hungarian invention, if not a "Hungaricum":

4 **processzor** 6558 >szerver ventillátora leállt. **Processzor** ventillátor (The server's fan has stopped. Processor fan)<
3 **proci** 46561 >gépben **proci** ventillátor rossz (processor's fan is out of work in the computer)<
6 **vincsit** 39885 a megfelelő **vincsit** (the right HDD)<
127 **vinyót** 49989 >nem ismeri fel a **vinyót** (won't recognize the HDD)<

4. Conclusions

What was relevant to examine from the whole corpus – the use of the language of information technology, especially the ESP terms of it – represents only a small

proportion of the text. So what makes IT language? Surely it is not only the terminology that makes it. In the framework of this research I restricted myself to the lexicographical investigation of the text with some attention to the syntactics and morphology of the sentences, not going into textological aspects. The genre of the Help Desk documents was not to be expected to be too technical or specific. People with presumably – and really – little knowledge of information technology informed Help Desk workers about problems they had to face with mainly in their offices. No wonder that the most frequently used word was nyomtató, the printer. Help Desk workers are experts, but the measures they had to undertake were restricted to the narrow scope of the problems they had to react to.

Nevertheless, the language that was used in the document fully represents the IT language used at that segment of the Company.

In a later (or second) phase of the research, new data will have to be added to the present one, making the corpus a real monitor corpus. Attaching the results of the analysis of other subcorpora to the present one, we can obtain a unified corpus; it might then be possible to write software to improve the language use at the Company. This software will have to be upgraded from time to time with the monitor corpus.

Another and probably much more important outcome of the research could be the extension of the scope of research to the analysis of different ESP *teaching materials*. The quality of the textbooks is of utmost importance as they are the very ones from which the pupils, and later the students, learn about informatics. Special attention must therefore be paid to the language use of them, as well as the mass media (television, radio, newspapers, magazines etc) which also have great influence on the students. As it was suggested in the introduction, a controlled dictionary or encyclopaedia on the Internet, possibly with an Internet forum, selected links to other websites with special content, etc., could be an invaluable help in eliminating the chaos and in promoting the proper language use. By establishing a website on the Internet, aiming for the controlled and unified use of IT terms, and involving linguists, IT experts, and eager volunteers, a fast reaction to the new and new challenges could be reached. If experts with a high reputation and prestige were part of the project, the acceptance of the terms would probably be general and – in my view – an ongoing operation of such a website would be very efficient. Moreover, the website would form the corresponding part of the Hungarian scientific language as well as the language use of the broader public (students, teachers, readers, viewers and listeners, etc).

The results of the research clearly show, and conclusively prove, that the use of IT language terminology is highly mixed, and the overuse of English words, with or without Hungarian affixes, cause major morphological problems – this kind of language (ab)use probably can described as "Hunglish" or at least "Englarian". Although it surely creates various problems and difficulties everywhere where information technology is concerned, those in the education of informatics seem to be the most serious, rippling around as students pass all the confused and Hunglish terms they once learned. To avoid this, as we have emphasized before, we should promote an effective (and probably Internet-based) co-operation between linguists, IT experts, students, teachers, and researchers.

Acknowledgements

Special thanks to the company of Morphologic for their invaluable assistance in this research.

References

[1] István Boda and Judit Károly-Porkoláb, Metaforák a kognitív nyelvészetben, Informatikai szaknyelvi metaforák vizsgálata, *Szemiotikai szövegtan* **13** (2000), 73–85.

[2] N. Chomsky, *Aspects of the theory of syntax*, MA: MIT Press, Cambridge, 1965.

[3] N. Chomsky, *Language and mind*, Harcourt Brace, New York, 1968.

[4] Magyar Számítógépes Nyelvészeti Konferencia, (Dr. Alexin Zoltán and Csendes Dóra, eds.), Szegedi Tudományegyetem, Szeged, Egyetemi Nyomda, 2003.

[5] Special Issue on Using Corpora in Language Teaching and Learning, *Language Learning & Technology* **5**, no. 3, (C. Tribble and M. Barlow, eds.) (September 2001), `http://llt.msu.edu/vol5num3`.

[6] T. McEnery and A. Wilson, *Corpus linguistics*, Edinburgh University Press, Edinburgh, 1996.

[7] The Corpus Research Group, University of Birmingham. Many useful links, including access to an email-based part-of-speech tagging service Michael Barlow's corpus linguistics site. Many useful links and sources of information. Covers a wide variety of languages: `http://www.ruf.rice.edu/~barlow/corpus.html`.

[8] University Centre for Computer Corpus Research on Language (UCREL), University of Lancaster. Many useful links and a Web-based part-of-speech tagging service: `http://www.comp.lancs.ac.uk/computing/research/ucrel`, other Web site: `http://www.scp.com`.

ATTILA KÁLMÁN JUHÁSZ
LANGUAGES DEPARTMENT
COLLEGE OF DUNAÚJVÁROS
KALLÓS D. U. 1.
H–2400 DUNAÚJVÁROS
HUNGARY

E-mail: juhasza@mail.duf.hu

(Received April, 2005)