

Evaluation of terminology extraction tools

**TExt for TWIN
System Quirk
Xerox TermFinder**

Britta Heidemann, Martin Volk

**Universität Zürich
Institut für Informatik
Abteilung Computerlinguistik**

12. Dezember 1999

Contents

Definition.....	3
Description of the systems	4
Text for TWIN.....	4
System Quirk.....	5
Xerox TermFinder.....	6
Evaluation	7
Procedure.....	7
Categories and Patterns	8
Decisions in special cases.....	9
Results 1 – The extracted items	10
Number of items extracted by each system	10
Classification of all extracted items.....	10
Further Observations	12
Results 2 – Classified terms.....	15
Number of terms according to category	15
Percentage of terms according to category.....	16
Percentage of terms within each category	20
Results 3 - Recall.....	21
Conclusions	22
Precision and Recall	22
Comparison of search methods	22
Which system for which task?.....	23

Appendices

Definition: Term

A term is a word that has a certain meaning in a certain subject area which differs from the general meaning (in the subject area Computer Science, the following are terms: *windows*, *user*, *account*, *site*, *session* or verbs like *connect*, *disconnect*).

Furthermore, a term is a word which is used only in a certain subject area (*UNIX computer*, *SUID program*).

Further criteria: The specific meaning of the term is part of the shared knowledge of author and reader – if the reader does not know the term, text understanding is impossible. Therefore, Unix commands such as *ls*, *pwd* etc. have to be regarded as terms: *If you give the pathname to a directory, ls will list the directory but it will not change your working directory. The **pwd** command in the following example shows that.*

Description of the systems

TExt for TWIN

TWIN (Termbase for Windows):

TWIN is a “multi-user terminology database system developed under Microsoft Access 2.0”, “replicated database with management features to administer virtual terminology networks”.¹

Term extraction tool (TExt) for TWIN:

“This tool will extract terminology from an input text file, compare it with a TWIN-termbase, and, depending on the selected output options, paste target language equivalents into the text or create a bilingual glossary, or a list of terms not found in the termbase. These features also work with multiword terms. Supported input language is English. Supported input format: Word for Windows and HTML. Supported editor: Word for Windows. Other features, as, for example, support of other input languages or editors may be implemented on a project basis.”²

TExt Solution to frequent problems in term extraction³

Combination of Procedures:

- ⇒ run statistical text analysis on normalised input
- ⇒ run multi-level linguistic filter on statistical results

Analysis Steps

- ⇒ Text analysis and normalisation (tagging and lemmatisation)
- ⇒ Multiword Boundary detection
 - ⇒ Criteria for Multiword boundaries: Formal criteria (sentence/subsentence boundary); Linguistic criteria (e.g. multiword sequences have no modal verb form).
- ⇒ Statistical Analysis of Multiword candidates
- ⇒ Linguistic Filtering of Multiword candidates
 - ⇒ first level filter (legal multiword category sequences)
 - ⇒ second level filter (exclude stop words of a given class: adjectives, prepositions; exclude certain patterns)
- ⇒ Output Generation

Hardware: Pentium-PC, 16 MB RAM, “Fast Hard Disk”⁴.

Price (TExt for TWIN): US\$ 1800.

Company: triS System Solutions & Services GmbH & Co., 81929 München/ Deutschland

¹ <http://www.tri-s.de/twinv11.html>

² <http://www.tri-s.de/tools.html>

³ Gr. Thurmair triS/GMS: Term Extraction from Texts.

⁴ <http://www.tri-s.de/twinv11.html>

System Quirk

“System Quirk is a flexible package of integrated tools for building and managing term bases”.⁵

KonText (Knowledge on Text): “...is a text analysis program that allows you to select one or more text files from a corpus, and to view, edit, print or analyse the text file(s) by performing tasks such as Key Word in Kontext (KWIC), Weirdness, wordlists and indexes within user defined constraints.”

Weirdness: “The task Weirdness creates a statistical approach to terminology extraction from text by comparing the relative frequencies of words that occur in specialist texts with their relative frequencies in general language corpora.”

Ferret: “Ferret is a text analysis program that allows you to analyse a number of texts specifically for compound terms. It makes use of stoplists to create term boundaries and also allows you to create/remove lists of known terms for future analysis and also for use withing other applications such as KonText. Term lists are created by considering those compounds that appear between words in the stoplists.”

ColloQuator: “A collocation statistics application. Provides statistical evidence for collocations. The full version of this application can be seeded with a number of tokens/terms and will return statistical information about terminess for compounds.”

Company: School of EEITM, University of Surrey, Grossbritannien

Hardware: There are two versions of System Quirk available - the research (Unix) version and the PC version for Windows 95/98/NT (Minimum Pentium 166MHz processor, 10MB disk space, 32MB RAM)

Price of the PC version: £ 109 (Academic Institutions) resp. £ 219 (Commercial Institutions).

⁵ <http://www.computing.surrey.ac.uk/ai/SystemQ/links2.html>

Xerox TermFinder

TermFinder, one part of the Xerox Terminology Suite, is supposed to help collecting the data then managed by the Term Organizer. Other components of the system are Term Checker (consistency check) and Xerox Multilingual Assistant (translation of words in context).

“This tool builds terminology databases from existing texts and documents. It extracts terms and provides a list that can be verified afterwards by a terminologist. It can be used with a single language or two, providing the equivalent translation and word in the target language for a source-language term. Using Xerox TermFinder with Xerox TermOrganizer results in large productivity gains and ensures accuracy and consistency in documents and aid in the creation of customised glossaries.

Xerox Term Finder doesn't have any direct key competitor. Only Xerox achieves bilingual linguistic multiword term extraction: no other terminology tool offers fully integrated source text analysis and automatic extraction tools to identify new terms. Tests at XRCE suggest Xerox's tool is eight time faster and cheaper than traditional manual methods of building terminology.”⁶

Sentences of the source and target documents are aligned automatically and can be checked manually. The terminology data base is collected via monolingual or bilingual term extraction.

“Xerox Term Finder uses XeLDA, the unique Xerox linguistic engine. XeLDA is the result of twenty years of research in linguistics at the Palo Alto Research Center (PARC) and at the Xerox Research Centre Europe (XRCE). XeLDA uses very sophisticated computer entities called “transducers” that embed linguistic information in a very efficient, compact, language-independent way. XeLDA enables the Xerox TermFinder engine to quickly align sentences and find terms in Documents.”

“Built on top of Open Database Connectivity (ODBC), Microsoft`s database independent layer, Xerox TermFinder is independent of any specific database. Xerox TermFinder supports SGML, HTML,XML, iso-8859-1 and RTF documents.”

Hardware: Pentium-II-PC, 96MB RAM, 100 MB hard disk.

Price: ~ 20.000 \$ (for the complete Xerox Multilingual Suite)

Company: Xerox Multilingual & Knowledge Management Services, Meylan/France.

⁶ <http://www.mkms.xerox.com>

Evaluation

Procedure

1. Automatic extraction of terms from selected chapters of two UNIX books⁷ by System Quirk (Unix Version), TExt/ Tris, TermFinder/ Xerox.

The texts were HTML-texts. Ignoring the HTML tags they contained 17082 words. Each system automatically extracted one or two lists with items regarded as possible terms: Whereas Xerox TermFinder provided one list, System Quirk's results from the KonText and Ferret programs (see the system description above for details) were in the form of two lists with single items resp. multiwords. For the TExt system, we evaluated the list with items found in the TWIN-termbase (collected in a bilingual glossary) as well as the list with extracted items that were not found in this termbase.

N.B.: We only evaluated the output (the extracted items) of the systems. We did not judge the user friendliness or the architecture of the three systems.

2. Classification of the extracted items according to a list of categories and patterns (see below for details on establishing these categories).

Beginning with the extracted items Xerox TermFinder listed, we manually worked through the lists and annotated a category marker to each item in the list. Each item that we regarded as term was marked with a +. When finished with the lists of all three systems, a Perl-script was used to count the terms and the number of items in each category. In this report you find the tables with the results, furthermore you may want to have a look at the lists with all terms each system found in the appendices to this report.

3. Computation of "precision and recall".

Precision: How many of the automatically extracted items can be regarded as terms? The number of manually marked terms was counted with the help of a Perl-script. We then calculated the percentage of terms within the lists of extracted items each system provided.

Recall: How many terms in the given texts have not been found by the system? In the source texts (the Unix book chapters) all those terms found by one or more of the three systems were automatically marked as terms. We then manually worked through the chapters and added markers to all other items that were terms in the context of this subject area. With a Perl-script the terms in the texts were compared to the list of terms each system had found.

⁷ Jerry Peek, Tim O'Reilly & Mike Loukides: UNIX Power Tools. Second Edition, August 1997 (1120 Seiten). Daniel Gilly & the Staff of O'Reilly & Associates: UNIX in a Nutshell: System V Edition. Second Edition, June 1992 (444 Seiten).

Categories and Patterns

1. Noun (N)
2. Noun + Noun (N2)
3. Noun + Noun + Noun (N3)
4. Nounphrase + Preposition (*of*) + Nounphrase (NON)
5. Adjectiv/Adverb (A)
6. Adjectiv (or Participle) + Noun (AN)
7. Adjectiv + Noun + Noun (AN2)
8. Verb (V)
9. Verb + Verb (V2)
10. Verb + Noun (VN)
11. Verb + Noun + Noun (VN2)
12. Abbreviations (K)
13. Abbreviation + Noun (KN)
14. Other Combinations of Words (C)
15. Unix Commands (B)
16. Command + Noun (BN)
17. Noun + Command (NB)
18. Keyboard Shortcuts (T)
19. Computercode (PC)
20. Webaddresses (W)
21. Filenames (D)
22. Names, Institutions (P)
23. Wrong segmentation (S)

The categories and patterns above were partly established in advance, partly during the evaluation. The categories containing word classes (nouns, adjectives and adverbs), had been defined beforehand, for example. Others - such as Unix commands, noun plus command, filenames etc. - became necessary in the evaluation process.

Decisions in special cases

Homographs: A word was assigned to the class occurring most often in the text. In the list provided by Xerox TermFinder, for example, *answer* was classified as noun (occurring 3 times in the text) and not as verb (1 time), but *look* as verb (5 times) and not as noun (2 times).

Verbs/Commands: Whenever a verb might be a Unix command, it was classified as such (e.g. *find*, *open*, *undelete* etc.) – even if it occurred within the text as verb and not as command. The decision was determined by the fact that in this subject area the verb does have a special meaning.

Capital letters: In general, words beginning with capital letters were not marked as terms. An exception was made for proper nouns (names of systems and programs etc.): Both *emacs* and *Emacs*, for example, were classified as term.

Plural/Singular: If there was no lemmatization routine involved in the term extraction process, often both plural and singular were extracted. System Quirk, for example, counted 16 times the multiword term *relative pathname*, five times the plural multiword *relative pathnames* (another example: *window system* (5 times) and *window systems* (3 times)). In general, the singular form was counted as a term. Due to this, some words were not counted as terms since they occurred only as plural form.

Results 1 – The extracted items

Number of items extracted by each system

	TExT	Quirk	Xerox
# of items	2237*	1843**	2070

*TExT found 113 words in its termbase (which contained 4800 entries) – the system produced a glossary containing these words and their respective translations in a bilingual list. In addition, the system regarded 2124 unknown items as probable terms.

**Of which 329 items are multiwords.

Classification of all extracted items

Category	TExT	Quirk	Xerox
a (adjective)	112	167	21
an (adj. + noun)	185	61	328
an2 (adj. + two nouns)	47	6	60
b (command)	56	57	83
bn* (command + noun)	2	5	5
c (other combination)	268	50	137
d (filename)	60	70	108
k (abbreviation)	82	66	60
kn* (abbrev. + noun)	11	11	12
n (noun)	278	537	475
n2 (two nouns)	173	136	213
n3 (three nouns)	21	14	36
nb (noun + command)	11	1	9
nk (noun + abbreviation)	7	2	4
non (prepositional phrase)	467	0	70
p (name)	28	38	13
pc (computercode)	66	148	90
s (segmentation error)	219	54	209
t (keyboard shortcut)	7	2	8
v (verb)	84	403	45
v2 (two verbs)	0	0	3
vb (verb + command)	0	0	6
vn (verb + noun)	42	12	51
vn2 (two verbs)	4	2	17
w (webaddress)	2	0	3
extracted items	2237	1843⁸	2070

⁸ number of evaluated terms (down to Frequency Ratio 1).

Comments:

Category “other combinations of words” (c): In the TExt system, a high number of verb phrases were classified as c, which is the reason for the gap between TExt and the two other systems in this category.

Category “prepositional phrases” (non): While Xerox` TermFinder extracted only those prepositional phrases containing the preposition *of*, TExt also recognized phrases containing *from*, *with*, *at* etc. as possible terms.

Category “verb” (v): The words extracted by System Quirk contained a high number of participles, which were classified as verbs, due to missing context. As the lists extracted by the two other systems are concerned, those participles probably are part of the multiwords classified as an (adjective or participle plus noun).

Wrong segmentation (s): System Quirk produced comparably less segmentation errors than the two other systems. Both Xerox TermFinder and TExt seem to have problems with HTML tags within the texts: Often two words were directly attached to each other in the list of extracted items that were in the text separated by an HTML tag – obviously those tags should be replaced by a space to ensure proper segmentation and tokenization of the text.

Further Observations

Xerox TermFinder

Lemmatization sometimes produced unwelcome results. For instance, the system mistakenly performed the following reductions:

operating system => operate system

or

stopped jobs => stop jobs

TermFinder delivered a high number of multiword combinations that contained general purpose adjectives such as *different, good, last, many, new, old, same, simple, special* plus noun (phrase):

simple alias

simple command

simple command line

simple design

simple method

simple pipeline

simple type error

old command

old content

old directory entry

old version

Such combinations are most likely no terms. They should be sorted out with the help of a stop list.

TExt for TWIN

The system often produced wrong word forms due to lemmatization of adverbial participles:

line-oriented => *line-orient*

berkeley-based (mechanism) => *berkeley-base*

(This is not true in all instances: While Xerox' TermFinder counted nine times the phrase *delete file*, TExt counted eleven times *deleted file*.)

For the TExt system, category *c* (other combinations of words) contains more verb phrases than the same category for the other two systems. For example:

performing non-sequential output

plan to use into external variables

print character after character

As with the Xerox system, there were a high number of multiword combinations that contained adjectives such as *common*, *different*, *new*, *old*, *own*, *simple* plus noun. Again, these are no likely terms:

common beginner

common control character

common login name

common program

common type

new application

new directory

new entry

new file

new group

new name

The TExt system extracted a higher number of prepositional phrases than the other systems containing *from*, *with*, *at* etc., e.g.:

new directory with its absolute pathname

password for this group

pathname to the play directory

private shell file in each project

There seems to be no lemmatization routine involved in the collection of terms for the glossary list: Evidence for this assumption is the high number of different verb forms and plural nouns in this list. (Plural nouns were not marked as terms!)

Furthermore, the translations found in the glossary often were wrong:

PC => *pC* – *parallel C*

Sometimes homonyms were translated as belonging to the “wrong” (unusual) word class in this subject area:

links => *verknüpfen*

loop => *schleifen*

Both words were manually classified as nouns by the evaluator.

System Quirk

The single word list was sorted according to the so-called “frequency ratio”, which compares the frequency of a word within a general corpus with its frequency of occurrence in a special subject area. Those words that have a high “frequency ratio” belong to the specialized vocabulary. Words with a low frequency ratio are not supposed to be terms (except for those words with a special meaning parallel to the general one such as *window*, *system*, *network*), therefore we cut the list of possible terms below frequency ratio = 1. That means: Only 1517 of 2278 extracted single word items have been classified, since we did not assume to find any terms within those with a frequency ratio less than 1.

The list contained a number of words that could not be determined with respect to our categories (e.g. conjunctions and prepositions such as *via*, *below*, *if...*), so they were assigned to category c (other combinations of words). Pronouns were classified as n (Noun). The high number of participles were counted as verbs, therefore System Quirk has a high score in the category “verb” with respect to the two other systems. (There, those participles often occurred in noun phrases and therefore were marked as adjective (category an)).

The list with multiword combinations contained no verb or prepositional phrases, but only noun phrases. This is probably due to the usage of stop lists.

Results 2 – Classified terms

Number of terms according to category

category	TExt	Quirk	Xerox
a (adjective)	6	7	1
an (adj. + noun)	23	9	28
an2 (adj. + two nouns)	4	0	5
b (command)	21	31	35
d (filename)	1	4	4
k (abbreviation)	12	17	15
kn* (abbreviation + noun)	6	2	9
n (noun)	68	116	125
n2 (two nouns)	100	55	116
n3 (three nouns)	8	8	12
nb (noun + command)	4	1	5
nk (noun + abbreviation)	3	1	1
p (names)	0	2	0
pc (computercode)	0	3	0
non (prepositional phrase)	3	0	0
t (keyboard shortcut)	7	1	7
v (verb)	13	37	3
vn (verb + noun)	1	3	6
vn2 (verb + two nouns)	0	0	2
# of all terms	284	297	375
# of extracted items	2237	1843 ⁹	2070
percentage of terms within the list of extracted items	12,69%	16,11%	18,1%

The numbers in the table above show the distribution of categories within all manually classified terms of the extracted items.

For example, Xerox TermFinder extracted 2070 items. Out of these 375 were regarded as terms by the evaluator. That means that only 18,1% of all extracted items were terms. The numbers in the above table show the distribution of the terms over the categories. It can be seen that mostly nouns and noun phrases were regarded as terms.

The percentages at the bottom of the table show the so-called “precision” of each system in term extraction: How many of the automatically extracted items in the list are terms?

⁹ number of extracted terms (down to Frequency Ratio 1).

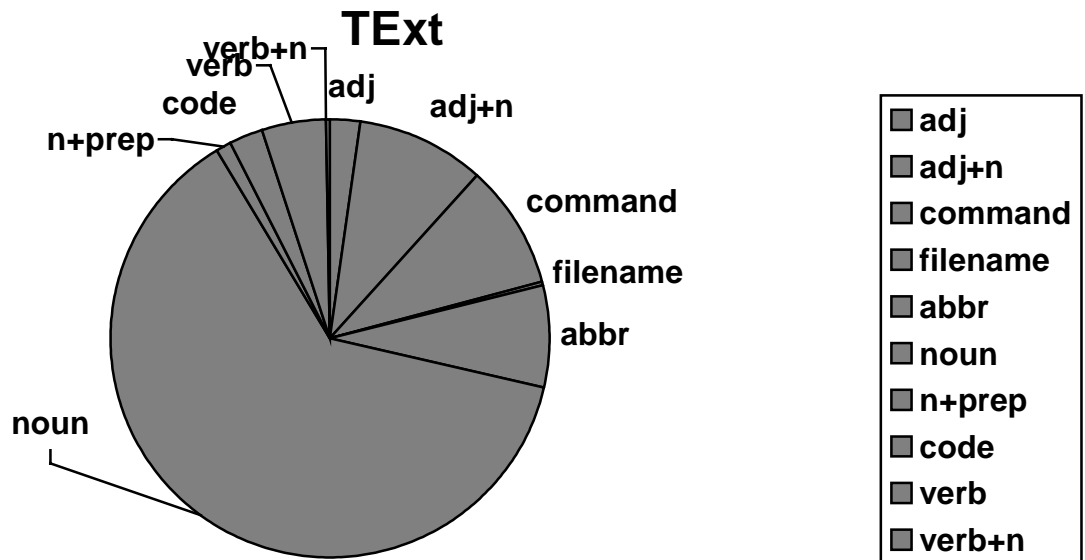
Percentage of terms according to category

NB: The category s is not shown since it does not contain any terms.

NB2: All numbers are percentages.

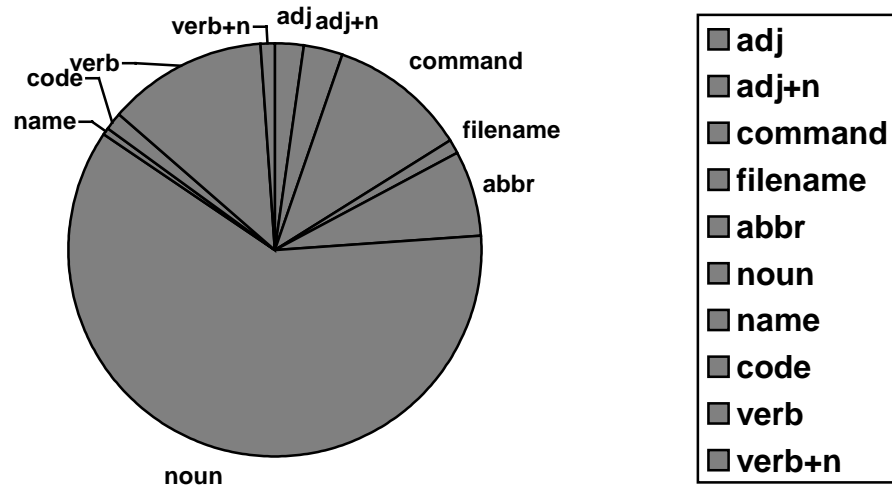
	TExt	Quirk	Xerox
a (adjective)	2,11	2,35	0,27
an/an2 (adj. + noun)	9,51	3,03	8,80
b/bn/nb/vb (command + combination)	8,80	10,77	10,67
d (filename)	0,35	1,01	1,07
k/kn/nk (abbreviation + combination)	7,39	6,73	6,77
n/n2/n3 (nouns)	61,97	60,27	67,47
non (prepositional phrase)	1,05	0	0
p (names)	0	0,67	0
pc/t/w (computer code, keyboard shortcut, webaddress)	2,46	1,34	1,86
v (verb)	4,58	12,46	0,80
vn/vn2 (verb+ noun)	0,35	1,01	2,13

The numbers in the table above show the distribution of categories within all manually classified terms in percentages. We have grouped similar categories in order to get a clear picture over general tendencies. For example, for Xerox TermFinder we have noted that 375 extracted items were regarded as terms. Out of these 375 terms 67,47% were nouns or noun sequences, 10,67% were command names or combinations of commands and nouns or verbs, 8,80% were adjective noun combinations and so on. On the following pages, you can see those numbers in a diagram for each system with further explanations on the program's performance.



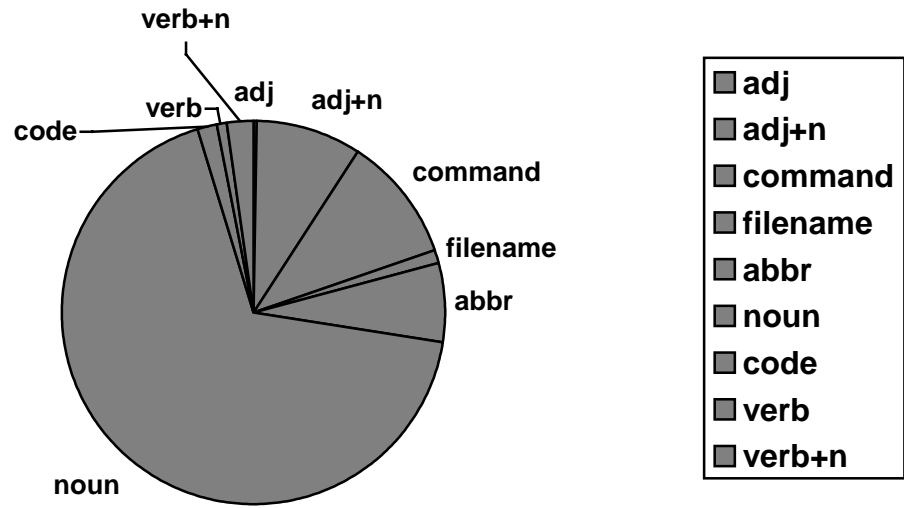
In the list of extracted terms provided by the TExt system, Unix commands or command plus noun combinations make up for 9,25% of all terms, followed by 8,37% adjectives and adjectives plus nouns.

Quirk



In addition to the nouns the terms provided by System Quirk consist to a high percentage of verbs (12,46%), followed by commands (plus nouns) (10,77%), and abbreviations (6,73%).

Xerox



The terms extracted by Xerox TermFinder consist of a high amount of nouns – and additionally of commands plus nouns (10,67%), adjectives plus nouns (8,80%), and abbreviations plus nouns (6,77%).

Percentage of terms within each category

The table below shows how many items in each category within the list of extracted items have been considered as terms by the evaluator. (I.e., how many of them have been marked with a +.) All numbers are percentages. For example, 37,13% of all nouns (n/n2/n3) that have been extracted by the TExt system were classified as terms, but only 5,35% of all adjectives.

NB: The categories s and p are not shown since they do not contain any terms.

	TExt	Quirk	Xerox
a (adjective)	5,35	4,19	4,76
an/an2 (adj. + noun)	11,58	13,43	8,52
b/bn/nb/vb (command + combination)	36,76	72,73	38,83
d (filename)	1,67	4,35	3,70
k/kn/nk (abbreviations + combination)	21,00	25,32	32,89
n/n2/n3 (noun)	37,13	26,06	34,94
non (prepositional phrase)	0,64	0	0
p (names)	0	5,26	0
pc/t/w (computer code, keyboard shortcut, webaddress)	9,33	2,67	6,93
v (verb)	15,47	9,18	6,67
vn/vn2 (verb+ noun)	2,17	21,43	11,76

Comments

Generally speaking, the percentage of terms is highest among expressions consisting of one or more nouns or combinations of commands or abbreviations and nouns (categories b, k, n). Noun phrases plus prepositional phrases (non) are obviously no reliable pool for terms – even in the high number of multiwords in the category “prepositional phrase” provided by the TExt system only 0,64% were terms.

Results 3 - Recall

The so-called “recall” answers the question of how many terms have not been found by the system in the given source text.

The numbers in the table below were collected via the following procedure: In the source texts (the Unix book chapters) all those terms found by one or more of the three systems were automatically marked with <t> at the beginning and </> at the end of each word or phrase. We then manually worked through the chapters and added markers to all other items that were terms in the context of this subject area. With a Perl-script the list of all terms in the texts was compared to the list of terms each system had found.

	TExt	Quirk	Xerox
# of all terms in the texts (types)		586	
# of found terms (types)	284	297	375
# of terms (types) not found	302	289	211
recall in %	48,46%	50,68%	63,99%

The total number of terms in the texts was 586 – if all 586 terms had been found by one of the systems, its recall would have been 100%. Neither of the systems reached this number. As can be seen in the table above, Xerox TermFinder by far reached the best recall score with 63,99%. Quirk and TExt reached comparable scores of 50,68% resp. 48,46%.

The list of terms not found by the TExt system contained multi- and single-word expressions of all categories, whereas System Quirk had problems finding multiword expressions. Often parts of the multiwords could be found in the list of extracted words (e.g. the abbreviation *ID*), but not the whole expression (e.g. *user ID*).

The contrary is true for Xerox` TermFinder: The list of not found terms contained many single-word expressions, most often verbs (such as *connect*, *quit*, *print*, *execute* etc). This is due to the fact that the system does not regard/count verbs as terms.

For more detailed information, please compare the lists of all terms in the source text and terms found by the three systems in the appendices.

Conclusions

Precision and Recall

As can be seen in the table above, Xerox TermFinder has reached the highest percentages of precision (number of terms in the extracted list) and recall (number of terms found with respect to possible number), followed by System Quirk and TExt.

	TExt	Quirk	Xerox
Precision	12,69%	16,11%	18,11%
Recall	48,46%	50,68%	63,99%

It should be noted that an evaluation of term extraction tools necessarily suffers from a range of subjectivity factors. First and foremost, it is not clearcut what counts as a term. Second it is difficult how to value morphological variations (plural forms) or spelling variations (upper case vs. lower case spelling). Therefore the figures should not be taken as absolute values but rather as indications of comparative merits. The figures indicate that the three evaluated systems fare rather similar with regard to precision. The Xerox TermFinder has slight advantages since it focuses on noun phrases which account for the majority of phrases (at least in the subject area of Computer Science). All systems produce more than 80% noise in automatic term extraction. We believe this could be reduced by half by subject specific filters that use statistical and linguistic clues. Surprisingly no system reached more than 70% recall. Too many specially structured terms were missed by all of them.

Comparison of extraction methods

	TExt	Quirk	Xerox
Use of a data base during the evaluation procedure	yes (=> creation of a glossary)	no	no
Lemmatization	yes (but not in the glossary)	no	yes
Frequency of extracted items	yes	yes	yes
Ranking of extracted items	no	yes	no
Verbphrases as terms	yes	no	?
Prepositional phrases as terms	yes	no	yes (but only those with <i>of</i>)
Context examples	yes	no	yes
Bilingual term extraction	no	no	yes

Which system for which task?

Xerox TermFinder got the best scores for precision and recall and is the only system that extracts bilingual term lists from bilingual corpora. And still lacks some important features: It does not “rank” the extracted words with respect to their “terminess”. If the user is especially interested in verb or prepositional phrases, he might be disappointed, since TermFinder only extracts special types of those phrases.

The advantage of System Quirk is its “ranking” feature – the user can rely on the fact that the extracted items with a high frequency ratio have a high “terminess”. On the other hand, System Quirk does not extract verb or prepositional phrases at all and does not provide any contextual examples for the items extracted.

Besides the use of a data base during extraction (and the collection of known terms in a bilingual glossary), the TExt system offers special features that make the system attractive even if the user considers the scores for precision and recall the system received. For example, the system extracts verb phrases and prepositional phrases including those with prepositions like *from*, *with*, *at* etc., uses a number of lemmatization features during term extraction and gives context examples for the extracted terms.

Appendix 1
List of all terms

.c
.cshrc
.exrc
.login
.o
.profile
.txt
ASCII
AT&T-based system
BSD
Berkeley UNIX
Berkeley group mechanism
Berkeley-based mechanism
Berkeley-based version
Bourne-type shell
C
C shell
C shell manual page
CD-ROM
CTRL
CTRL-C
CTRL-D
CTRL-H
CTRL-Q
CTRL-S
CTRL-U
Command-Line argument
DOS
DOS utility
Emacs
Emacs backup file
File Access Permission
Filesystem Permission
Filesystem permission bit
GNU version
I/O
ID
INFINITY 6
Linux
MS-DOS
Macintosh
Multitasking
NFS
PC
RFS
SGID
SGID bit
SGID file
SGID program
SUID
SUID bit
SUID file
SUID program
Shell Prompt
Solaris
UID
UNI
UNIX
UNIX Account

UNIX Environment
UNIX account
UNIX book
UNIX command
UNIX computer
UNIX computer system
UNIX directory
UNIX directory tree
UNIX environment
UNIX filesystem
UNIX filesystem tree
UNIX kernel
UNIX operating system
UNIX power tool
UNIX program
UNIX session
UNIX shell
UNIX shell prompt
UNIX system
UNIX user
UNIX version
UNIX wildcarding
VMS
VMS system
Vi
[BACKSPACE]
[CTRL-H]
[DELETE]
absolute pathname
access
access bit
access file
access mode
access permission
access privilege
access time
accidental file deletion
account
account name
account use
administrative tool
administrator
alias
align
amount of storage
application
argument
ascii
asterisk
authorized system user
automatic setup
backup
backup file
backup tape
bash
berkeley
berkeley group mechanism
berkeley mechanism
berkeley unix
binary
binary character
bit

block	deleted file
bourne shell	digit
browse	dir
bug	directory
byte	directory entry
c	directory level
cancel a command	directory location
capability	directory name
capability value	directory permission
cat	directory tree
cd	directory user
character	disconnect
chgrp	disk
child directory	disk block
chmod	disk drive
chown	disk partition
clear	display
code	display format
command	ditroff
command date	dos
command exit	drive
command file	driver
command interpreter	dump
command line	edit
command logout	editor
command name	electronic mail
command ps -aux	emacs
command tool	emacs editor
command-line argument	email
compile	end
compiles	enter
compound command	entomb
computer	entry
computer network	environment
computer system	environment variable
computing	erase
connect	erase character
control character	error message
control code	escape
copy file	escape sequence
cron	executable
crontab	executable file
crontabs	execute
cs	execute access
cs	execute permission
ctrl	executes
current directory	exit
current working directory	expression
cursor	expunge
daemon	external variable
data	family tree
data block	figure
data file	file
database	file access
default	file access permission
default group	file bar
default path	file content
default rm	file datum
default search path	file datum block
default system	file erasure
default umask	file group
default value	file name
delete	file owner

file ownership	line-oriented editor
file permission	link
file protection	linkname
file recovery system	linux
file removal	listing
file server	ln
file space	ln -s
file type	local computer
file user	login
file-deletion	login file
file-deletion protection	login name
filename	login session
filename extension	login shell
filesystem	login time
filesystem check	logout
filesystem layout	loop
filesystem root	loophole
filesystem tree	lpr
filesystem type	ls
filter	lsdel
find	macintosh
flag	mail
folder	mailbox
foo	mailbox file
freeze	manpage
fsck	manpages
generic command	manual page
gnu	memory
gnu version	menu
group	microsoft
group ID	microsoft windows computer
group access	mkdir
group guest	mode
group id	mode bit
group membership	modem
group name	multi-task
group ownership	multi-tasking operating system
group-write permission	multitasking
hard link	multiuser
hardware	multiuser operate system
help	multiuser operating system
helv	mv
home directory	navigation
host	network
hostname	network communication
html	network connection
id	network filesystem
initialization	newgrp
inode	news
inode number	nfs
input	non-printing character
input/output	non-sequential output
insert	nroff
install	octal digit
interface	octal number
interpreter	od
interrupt character	online
kermit	open
kernel	operating system
keyboard	option
korn shell	output
ksh	override protection
library	owner

owner access	rm
owner permission	root
ownership	root directory
paging	root password
parent	screen
parent directory	screen editor
passwd	screen-oriented program
passwd file	script
password	scroll
password file	search path
path	security
path setting command	sequential output
pathname	server
pc	session
performance	set group ID
permission	set user ID
permission character	setenv
permission control	setgid
permission in a directory	setup
personal computer	setup file
personal computer operating system	sgid
pipe	sgid file
pipeline	sh
power tool	shell
print	shell file
print work directory	shell manual page
printer	shell prompt
process	shell script
program	shell setup file
program file	shell start
program name	shell-script
program output	shelltool
programming	shortcut
programming language	shutdown
project directory	single-user system
prompt	site
protection	software environment
ps	source code
pwd	source control
quit	source file
r	standard input
range	standard output
read	standard unix
read access	startup file
read and write access	startup time
read permission	sticky bit
read/write/execute permission	stopped job
readable	storage
recompile	store
recover	stored
recoverable	string
relative pathname	subdirectory
remote computer	subroutine
remote computer file	subshell
remote login	suid
remote system	suid file
remote-login	sun workstation
remote-login program	sunos
rename	superuser
reset	superuser access
resets	superuser password
restarts	switch
rlogin	symbolic link

system
system administrator
system call
system command
system file
system message
system path
system program
system to system
tar
tcsh
telnet
termcap
terminal
terminal database
terminal description
terminal driver
terminal emulation
terminal emulation program
terminal number
terminal screen
terminal type
terminate
terminfo
text editor
text file
textfile
th
tool
touch
trashcan
tree
tree structure
troff
tset
uid
umask
unalias
undelete
unix
unix computer
unix computer system
unix directory
unix directory tree
unix environment
unix filesystem
unix filesystem tree
unix kernel
unix program
unix programming environment
unix shell
unix system
unlink
user
user ID
user ID number
user directory
user entry
user interface
username
utility
utility program
uucp

variable
vi.
video display
virtual
virtual memory
virtual memory technique
vms
wildcard
wildcard match
wildcarded name
window
window system
work group
working directory
workstation
world write
world-write access
world-write permission
writable
write
write access
write permission
x
y

Appendix 2

List of all terms found by the 3 systems (The number indicates by how many of the systems a term was found)

1 .c	2 UNIX computer system
1 .cshrc	2 UNIX directory
1 .exrc	2 UNIX directory tree
1 .login	2 UNIX environment
1 .o	2 UNIX filesystem
1 .profile	2 UNIX kernel
1 .txt	1 UNIX power tool
1 ASCII	1 UNIX program
1 AT&T-based system	2 UNIX session
2 BSD	1 UNIX shell
1 Berkeley UNIX	1 UNIX shell prompt
2 Berkeley-based mechanism	2 UNIX system
2 Bourne-type shell	2 UNIX user
2 C	2 UNIX version
1 CD-ROM	1 UNIX wildcarding
2 CTRL	1 VMS
2 CTRL-C	2 VMS system
2 CTRL-D	1 Vi
2 CTRL-H	3 absolute pathname
2 CTRL-Q	2 access
2 CTRL-S	1 access bit
2 CTRL-U	1 access file
1 Command-Line argument	3 access mode
1 DOS	1 access privilege
2 DOS utility	1 access time
2 Emacs	2 accidental file deletion
1 File Access Permission	2 account
1 Filesystem permission bit	2 account name
1 GNU version	2 account use
2 I/O	3 administrator
2 ID	3 alias
1 INFINITY 6	1 align
2 Linux	1 amount of storage
1 MS-DOS	1 application
2 Macintosh	1 argument
1 Multitasking	1 ascii
2 NFS	3 asterisk
2 PC	1 authorized system user
2 RFS	1 automatic setup
2 SGID	3 backup
1 SGID bit	1 backup file
2 SGID file	3 backup tape
2 SGID program	2 bash
2 SUID	2 berkeley
1 SUID bit	2 berkeley group mechanism
2 SUID file	2 berkeley mechanism
2 SUID program	1 berkeley unix
1 Solaris	2 binary
1 UID	2 bit
2 UNIX	1 block
1 UNIX Account	1 bourne shell
1 UNIX Environment	3 browse
1 UNIX account	3 bug
1 UNIX book	2 byte
2 UNIX command	2 c
2 UNIX computer	2 capability
	1 capability value
	2 cat
	1 cd
	3 character
	3 chgrp
	2 child directory
	2 chmod

2	chown	2	display
1	clear	2	display format
2	code	1	ditroff
2	command	1	dos
1	command date	2	drive
2	command exit	1	driver
2	command file	2	dump
2	command interpreter	2	edit
2	command line	2	editor
2	command logout	1	electronic mail
2	command name	2	emacs
2	command ps -aux	1	emacs editor
2	command tool	3	email
1	command-line argument	1	end
1	compile	1	enter
1	compiles	2	entomb
2	compound command	2	entry
2	computer	2	environment
3	computer network	1	environment variable
3	computer system	2	erase
1	computing	1	erase character
1	connect	3	error message
2	control character	1	escape
2	control code	1	escape sequence
1	copy file	1	executable
2	cron	3	executable file
2	crontab	1	execute
1	crontabs	1	execute access
2	csh	1	executes
1	cshrc	3	exit
1	ctrl	2	expression
2	current directory	2	expunge
1	current working directory	1	external variable
3	cursor	3	family tree
3	daemon	2	figure
1	data	2	file
1	data block	2	file access
2	data file	1	file access permission
2	database	2	file bar
2	default	1	file content
3	default group	1	file datum
3	default path	1	file datum block
2	default rm	1	file erasure
2	default search path	1	file group
1	default system	1	file name
2	default umask	1	file owner
3	default value	1	file ownership
2	delete	2	file permission
2	deleted file	1	file protection
2	digit	2	file recovery system
3	dir	1	file removal
3	directory	1	file server
3	directory entry	1	file space
2	directory level	2	file type
3	directory location	1	file user
2	directory name	1	file-deletion
1	directory permission	2	file-deletion protection
3	directory tree	3	filename
2	directory user	1	filename extension
2	disconnect	3	filesystem
2	disk	2	filesystem check
2	disk drive	3	filesystem layout
2	disk partition	1	filesystem root

1	filesystem tree	2	ls
1	filesystem type	2	lsdel
1	filter	1	macintosh
2	find	3	mail
2	flag	2	mailbox
3	folder	2	mailbox file
3	foo	1	manpage
3	freeze	2	manpages
2	fsck	2	manual page
2	generic command	2	memory
1	gnu	2	menu
1	gnu version	1	microsoft
2	group	1	microsoft windows computer
2	group ID	2	mkdir
1	group access	2	mode
1	group guest	2	mode bit
1	group id	3	modem
1	group name	1	multi-task
3	group ownership	1	multi-tasking operating
1	group-write permission		system
1	hard link	1	multitasking
2	hardware	2	multiuser
1	help	1	multiuser operate system
1	helv	1	mv
1	home directory	1	navigation
2	host	2	network
3	hostname	1	network filesystem
1	html	3	newgrp
1	id	1	news
2	initialization	1	nfs
3	inode	1	non-printing character
1	inode number	2	nroff
2	input	3	octal digit
1	input/output	2	octal number
2	insert	2	od
1	install	1	online
2	interface	1	open
3	interpreter	1	operating system
1	interrupt character	2	option
2	kermit	2	output
3	kernel	2	override protection
2	keyboard	2	owner
1	korn shell	1	owner access
2	ksh	1	owner permission
2	library	2	ownership
1	line-oriented editor	2	paging
2	link	2	parent
3	linkname	3	parent directory
1	linux	2	passwd
3	listing	2	passwd file
2	ln	3	password
1	ln -s	1	password file
2	local computer	2	path
3	login	3	pathname
1	login file	1	pc
2	login name	2	performance
1	login session	2	permission
2	login shell	1	permission character
2	login time	1	permission control
3	logout	1	permission in a directory
3	loop	2	personal computer
1	loophole	1	pipe
2	lpr	1	pipeline

1 power tool	1 shell prompt
1 print	2 shell script
1 print work directory	3 shell setup file
2 printer	2 shell start
1 process	1 shell-script
2 program	3 shelltool
2 program file	2 shortcut
2 program name	2 shutdown
2 program output	2 single-user system
1 programming	1 site
1 programming language	2 source code
3 project directory	1 source file
2 prompt	3 standard input
2 protection	1 standard output
2 ps	1 standard unix
3 pwd	2 startup file
2 quit	1 sticky bit
2 r	1 stopped job
1 range	2 storage
1 read	1 store
1 read access	1 stored
1 read permission	1 string
2 readable	3 subdirectory
2 recompile	2 subroutine
1 recover	3 subshell
2 recoverable	1 suid
3 relative pathname	1 suid file
3 remote computer	1 sun workstation
1 remote computer file	1 sunos
1 remote login	3 superuser
3 remote system	1 superuser access
1 remote-login	3 superuser password
1 remote-login program	2 switch
2 rename	2 symbolic link
1 resets	2 system
1 restarts	3 system administrator
3 rlogin	1 system call
3 rm	2 system command
3 root	1 system file
3 root directory	2 system message
3 root password	2 system path
2 screen	2 system program
2 screen editor	1 system to system
2 screen-oriented program	3 tar
2 script	2 tcsh
3 scroll	3 telnet
2 search path	3 termcap
2 security	3 terminal
1 sequential output	1 terminal database
1 server	2 terminal description
3 session	1 terminal emulation
2 set group ID	2 terminal emulation program
2 set user ID	2 terminal number
1 setenv	2 terminal screen
1 setgid	2 terminal type
1 setup	1 terminate
3 setup file	2 terminfo
1 sgid	3 text editor
1 sgid file	3 text file
2 sh	3 textfile
3 shell	1 th
1 shell file	2 tool
2 shell manual page	3 touch

3 trashcan
3 tree
1 tree structure
1 troff
2 tset
1 uid
3 umask
1 unalias
3 undelete
1 unix
1 unix computer
1 unix computer system
1 unix directory
1 unix directory tree
1 unix environment
1 unix filesystem
1 unix filesystem tree
1 unix kernel
1 unix program
1 unix programming environment
1 unix shell
1 unix system
2 unlink
2 user
2 user ID
2 user ID number
1 user directory
1 user entry
1 user interface
3 username
2 utility
2 utility program
2 uucp
2 variable
1 vi.
2 video display
2 virtual
1 virtual memory
1 vms
2 wildcard
1 wildcard match
2 wildcarded name
3 window
3 window system
2 working directory
2 workstation
1 world-write access
1 world-write permission
3 writable
1 write
2 x
1 y

Appendix 3

All terms found by TExt

ASCII
BSD
Berkeley UNIX
Berkeley-based mechanism
Bourne-type shell
C
CTRL
CTRL-C
CTRL-D
CTRL-H
CTRL-Q
CTRL-S
CTRL-U
DOS utility
Emacs
File Access Permission
GNU version
I/O
ID
Linux
Macintosh
Multitasking
NFS
PC
RFS
SGID
SGID file
SGID program
SUID
SUID file
SUID program
Solaris
UID
UNIX
UNIX Account
UNIX book
UNIX command
UNIX computer
UNIX computer system
UNIX directory
UNIX directory tree
UNIX environment
UNIX filesystem
UNIX kernel
UNIX session
UNIX shell prompt
UNIX system
UNIX user
UNIX version
VMS
VMS system
Vi
absolute pathname
access mode
accidental file deletion
account name
account use
administrator
alias
amount of storage

asterisk
authorized system user
backup
backup file
backup tape
bash
binary
browse
bug
byte
cat
character
chgrp
child directory
chmod
clear
command exit
command file
command interpreter
command line
command logout
command name
command ps -aux
command tool
command-line argument
compound command
computer network
computer system
control character
control code
cron
crontab
current directory
cursor
daemon
data block
data file
default group
default path
default rm
default search path
default value
delete
deleted file
dir
directory
directory entry
directory level
directory location
directory name
directory tree
directory user
disconnect
disk drive
disk partition
display format
edit
email
entomb
environment variable
erase
error message
escape

executable file	passwd
exit	passwd file
expression	password
expunge	pathname
family tree	performance
file bar	permission in a directory
file permission	personal computer
file recovery system	program file
file space	program name
file type	program output
file-deletion	programming language
file-deletion protection	project directory
filename	prompt
filesystem	pwd
filesystem check	quit
filesystem layout	readable
filesystem tree	recompile
folder	recover
foo	recoverable
freeze	relative pathname
generic command	remote computer
group ID	remote system
group guest	remote-login
group ownership	rename
hostname	rlogin
initialization	rm
inode	root
inode number	root directory
insert	root password
interpreter	screen editor
interrupt character	screen-oriented program
kermit	scroll
kernel	search path
linkname	session
listing	set group ID
local computer	set user ID
login	setup file
login file	shell
login name	shell manual page
login shell	shell prompt
login time	shell setup file
logout	shell start
loop	shell-script
mail	shelltool
mailbox	shortcut
mailbox file	shutdown
manpage	single-user system
manual page	source code
menu	source file
mkdir	standard input
mode bit	stopped job
modem	subdirectory
multi-task	subroutine
multi-tasking operating system	subshell
multiuser	superuser
newgrp	superuser password
nroff	symbolic link
octal digit	system administrator
octal number	system call
od	system command
operating system	system message
override protection	system program
parent directory	system to system

tar
telnet
termcap
terminal
terminal description
terminal emulation
terminal emulation program
terminal number
terminal screen
terminal type
text editor
text file
textfile
touch
trashcan
tree
tset
umask
undelete
unlink
user ID
user ID number
user interface
username
utility program
video display
virtual
virtual memory
wildcard match
wildcarded name
window
window system
working directory
workstation
writable

Appendix 4
All terms found by Quirk

absolute pathname
access
access mode
account
administrator
alias
align
ascii
asterisk
automatic setup
backup
backup tape
bash
berkeley
berkeley group mechanism
berkeley mechanism
berkeley unix
binary
bit
bourne shell
browse
bug
c
capability
cat
cd
character
chgrp
chmod
chown
code
command
compile
compiles
computer
computer network
computer system
computing
connect
cron
crontab
csh
cshrc
ctrl
cursor
daemon
data
data file
database
default
default group
default path
default system
default umask
default value
delete
deleted file
digit
dir

directory
directory entry
directory location
directory tree
disconnect
disk
display
ditroff
dos
drive
dump
edit
editor
emacs
emacs editor
email
enter
entomb
entry
environment
erase
erase character
error message
executable
executable file
execute
execute access
executes
exit
expunge
family tree
figure
file
file access
file erasure
file ownership
file recovery system
file removal
filename
filesystem
filesystem layout
filesystem root
find
flag
folder
foo
freeze
fsck
gnu
group
group access
group id
group ownership
hardware
helv
home directory
host
hostname
html
id
initialization
inode
input

insert	programming
install	project directory
interface	prompt
interpreter	protection
kermit	ps
kernel	pwd
keyboard	quit
korn shell	r
ksh	read
library	readable
link	recompile
linkname	recoverable
linux	relative pathname
listing	remote computer
ln	remote system
login	rename
login shell	resets
logout	restarts
loop	rlogin
loophole	rm
lpr	root
ls	root directory
lsdel	root password
macintosh	screen
mail	script
mailbox	scroll
manpages	security
memory	session
menu	setgid
microsoft	setup
microsoft windows computer	setup file
mkdir	sgid
mode	sgid file
modem	sh
multitasking	shell
multiuser	shell file
mv	shell script
navigation	shell setup file
network	shelltool
newgrp	shortcut
nfs	shutdown
nroff	standard input
octal digit	standard unix
od	startup file
online	storage
option	store
output	stored
override protection	subdirectory
owner	subroutine
ownership	subshell
paging	suid
parent	suid file
parent directory	sun workstation
passwd	sunos
password	superuser
password file	superuser access
path	superuser password
pathname	switch
pc	system
permission	system administrator
print	system path
printer	tar
program	tcsh

telnet
termcap
terminal
terminate
terminfo
text editor
text file
textfile
th
tool
touch
trashcan
tree
tree structure
troff
tset
uid
umask
unalias
undelete
unix
unix computer
unix computer system
unix directory
unix directory tree
unix environment
unix filesystem
unix filesystem tree
unix kernel
unix program
unix programming environment
unix shell
unix system
unlink
user
username
utility
uucp
variable
virtual
vms
wildcard
window
window system
workstation
writable
write
x

Appendix 5

All terms found by Xerox

.c
.cshrc
.exrc
.login
.o
.profile
.txt
AT&T-based system
BSD
Berkeley-based mechanism
Bourne-type shell
C
CD-ROM
CTRL
CTRL-C
CTRL-D
CTRL-H
CTRL-Q
CTRL-S
CTRL-U
Command-Line argument
DOS
DOS utility
Emacs
Filesystem permission bit
I/O
ID
Linux
MS-DOS
Macintosh
NFS
PC
RFS
SGID
SGID bit
SGID file
SGID program
SUID
SUID bit
SUID file
SUID program
UNIX
UNIX Environment
UNIX account
UNIX command
UNIX computer
UNIX computer system
UNIX directory
UNIX directory tree
UNIX environment
UNIX filesystem
UNIX kernel
UNIX power tool
UNIX program
UNIX session
UNIX shell
UNIX system
UNIX user
UNIX version
UNIX wildcarding

VMS system
absolute pathname
access
access bit
access file
access mode
access privilege
access time
accidental file deletion
account
account name
account use
administrator
alias
application
argument
asterisk
backup
backup tape
berkeley
berkeley group mechanism
berkeley mechanism
bit
block
browse
bug
byte
c
capability
capability value
character
chgrp
child directory
chown
code
command
command date
command exit
command file
command interpreter
command line
command logout
command name
command ps -aux
command tool
compound command
computer
computer network
computer system
control character
control code
copy file
crontabs
csh
current directory
current working directory
cursor
daemon
database
default
default group
default path
default rm

default search path	freeze
default umask	fsck
default value	generic command
digit	gnu version
dir	group
directory	group ID
directory entry	group name
directory level	group ownership
directory location	group-write permission
directory name	hard link
directory permission	hardware
directory tree	help
directory user	host
disk	hostname
disk drive	inode
disk partition	input
display	input/output
display format	interface
drive	interpreter
driver	kernel
dump	keyboard
editor	ksh
electronic mail	library
emacs	line-oriented editor
email	link
end	linkname
entry	listing
environment	ln
error message	ln -s
escape sequence	local computer
executable file	login
exit	login name
expression	login session
external variable	login time
family tree	logout
figure	loop
file	lpr
file access	ls
file access permission	lsdel
file bar	mail
file content	mailbox file
file datum	manpages
file datum block	manual page
file group	memory
file name	mode
file owner	mode bit
file permission	modem
file protection	multiuser operate system
file server	network
file type	network filesystem
file user	newgrp
file-deletion protection	news
filename	non-printing character
filename extension	octal digit
filesystem	octal number
filesystem check	open
filesystem layout	option
filesystem type	output
filter	owner
find	owner access
flag	owner permission
folder	ownership
foo	paging

parent	source code
parent directory	standard input
passwd file	standard output
password	startup file
path	sticky bit
pathname	storage
performance	string
permission	subdirectory
permission character	subshell
permission control	superuser
personal computer	superuser password
pipe	switch
pipeline	symbolic link
power tool	system
print work directory	system administrator
printer	system command
process	system file
program	system message
program file	system path
program name	system program
program output	tar
project directory	tcsh
protection	telnet
ps	termcap
pwd	terminal
r	terminal database
range	terminal description
read access	terminal emulation program
read permission	terminal number
relative pathname	terminal screen
remote computer	terminal type
remote computer file	terminfo
remote login	text editor
remote system	text file
remote-login program	textfile
rlogin	tool
rm	touch
root	trashcan
root directory	tree
root password	umask
screen	undelete
screen editor	user
screen-oriented program	user ID
script	user ID number
scroll	user directory
search path	user entry
security	username
sequential output	utility
server	utility program
session	uucp
set group ID	variable
set user ID	vi.
setenv	video display
setup file	wildcard
sh	wildcarded name
shell	window
shell manual page	window system
shell script	working directory
shell setup file	world-write access
shell start	world-write permission
shelltool	writable
single-user system	x
site	y

Bilingual terminology extraction plays an important role in the application of bilingual dictionary compilation, bilingual Ontology construction, machine translation and cross-language information retrieval etc. This paper addresses the issues of monolingual terminology extraction and bilingual term alignment based on multi-level termhood. Design/methodology/approach: A method based on multi-level termhood is proposed. The new method computes the termhood of the terminology candidate as well as the sentence that includes the terminology by the comparison of the corpus. • List of terms extracted from ST Extraction • List of terms to validate (accept or reject). • List is added to a termbase Translation • List is translated and additional data added. 1100. Monolingual extraction & validation. Importing & looking for equivalents. • List approved by a person in charge of terminology • When the client has requested there is an additional Approval step for client approval. 11. • EXTRACT-ALIGN 1ST step: monolingual terminology extraction in both languages. 2nd step: cross-linguistic matching using word-alignment or co-occurrence statistics to find equivalents. Commercial systems in this approach. • Concordance programs: WordSmith Tools, (terminology extraction from TMX). AntConc (free), • 3. Evaluation. Bate under evaluation. Sketch Engine. 19. Documentary Search Tools. Term-Extraction Tools. Terminology Research Tools. Data-Recording Tools. Electronic-Publishing Tools. • Evaluation of Terms and their Relationships. 28. Language Management and Terminology Harmonization 30. • Principal Tools for Terminology Work. Any terminological activity, from term identification through product delivery, can be performed manually. However, computerization allows unprecedented improvements in productivity, quality and accessibility.