LEXICAL PROFILES IN LITERATURE

The Influence of

Alice in Wonderland

and

Through the Looking Glass

by Lewis Carroll

on Present Day American and British English

SE: English Corpus Linguistics
Prof Kettemann, SS 2000

Helmut W. Klug
9212780

INTRODUCTION

The aim of this paper is to show the influence of Lewis Carroll's books *Alice in Wonderland* and *Through the Looking Glass* on present day American and British English. This should be achieved by comparing the corpus of the fictional text to several readymade corpora, as the LOB, Brown or the BNC. By comparing the wordlists I want to specify on certain keywords. By using these keywords I am going to compute kwic-lines from two different, specialised corpora: the corpora of *The Guardian* (the collection dating from 1990 to 1993), and the *Time Magazine* (the collection dating from 1989 to 1993). The key words will then be analysed within the context according to the following aspects: in which way are the words and /or phrases used today, has there been a shift, widening, … of meaning, and finally in which context are the keywords used today?

The structure of this paper is as follows: the first part is a theoretical approach towards corpora and corpus creation in which I want to give a basic approach to this topic by combining fundamental principles with up to date facts. The second part will be a summary of the practical work in which I am going to analyse different corpora as described above with the help of Mike Scott's *WordSmith Tools*, Version 3.00.

1. CORPUS CREATION

1.1. INTRODUCTION

The first step in doing corpus linguistic successfully is to decide on which corpus should be taken: ready-made corpora (LOB, FLOB, Brown, Frown, BNC, ...) or a corpus designed for a specific purpose. In this chapter I am going to give an introduction to the process of corpus creation (cf. Sinclair, 1991:13-26) to point out the various factors to be considered and on the other hand to provide the theoretical basis on which I made my choice of texts.

The first step to be taken is the decision upon what ought to be in the corpus and how the final selection should be organised. Before answering these questions it should be determined who collects a corpus. In the past decades and even nowadays, with corpus linguistics widely accepted and supported, it has been the linguist who builds the corpus. In the end it is the users and critics who have to decide on the constitution and balance of the corpus by analysing which texts the corpus provides. According to Sinclair (1991:13) the ideal way would be the sharing the work by: language-orientated social scientists creating corpora who decide on the types and proportions of the material used and linguists ready to analyse and describe any instance of language. Today corpus builders are aware of the social aspects of corpus analysis and when for example working with the British National Corpus (BNC) users can extract much more than simple concordances: subject-

field and medium of the texts, or age, gender and social status of the speaker or writer. But it is still the linguist who is most interested in corpus work.

When deciding on the content of the corpus the first consideration is the aim of the corpus which can range from giving a general sample of a language to picturing specific fields of language. The most common and most desired is a general corpus, which can be taken as a standard reference of a language:

> "We are therefore interested in creating a corpus which is maximally representative of the variety under examination, that is, which provides us with an as accurate a picture as possible of the tendencies of that variety, as well as their proportions. What we are looking for is a broad range of authors and genres which, when taken together, may be considered to 'average out' and provide a reasonably accurate picture of the entire language population in which we are interested." (McEnery and Wilson, 2samp.htm)

The reason for this can be found in the demands of the dominating area of corpus linguistics: the documentation of language and building of dictionaries and grammars. How important corpus linguistics has become for this work can be seen in the fact that in the year 1995 four monolingual dictionaries (Longman Dictionary of Contemporary English [3rd ed], Oxford Advanced Learner's Dictionary [5th ed], Collins Cobuild English Dictionary [2nd ed], and Cambridge International English Dictionary) were published almost simultaneously. (Rundell) In recent years specialised corpora have become increasingly common, ranging from specific domains like cancer research or academic texts (RAT Corpus) to the area of learner corpora used for ELT purposes.

1.2. Different Kinds of Corpora

Before discussing the process of corpus creation in detail I want to explain the main differences between already existing corpora. The main distinction is size, either of texts or the corpus as a whole and resulting from this corpora are divided into sample and monitor corpora.

The first corpus to be created in the 1960s, the Brown corpus, is a typical example of a sample corpus – it is a sample of written American English – and its creation has influenced corpus builders up to the 1980s. The main features of a corpus of this kind are the following: a classification into genres of text, a large number of short extracts, and a close to random selection of extracts within the different genres. An additional feature of some sample corpora is the tagging of the text where each word is coded according to its form and function. The main advantage of this kind of corpus is that a great amount of statistical information can be drawn from its analysis.

This has worked well as long as the main interest of researchers was the manipulation of the information gathered from corpora. Nowadays the main interest has shifted to interpreting corpus data, and now the limitations of these corpora are beginning to show. The main problem here is the size of the rather small sample texts (2000 words). The size of the whole corpus may be sufficient to portray general patterns of a language but it is not possible for researchers to do more detailed

analysis: a study of general text patterns will certainly produce an inappropriate result, since the sample texts are too brief.

The progress of hardware and software development has changed this situation. In the mid 1980s a new kind of corpus has been developed: one which has no final extent, like language itself. The main advantage of such a corpus is that it not only adds a historical dimension to corpus analysis, e.g. to portray the evolution of language, it also allows qualitative analysis of language, e.g. grammar, text patterns. What is more, it is able to portray the state of the language. Due to its size, and the continuously renewed texts researchers are at any time able to give an up-to-date picture of language.

1.3. Problems of Corpus Creation

Once the corpus builder has decided on the general content and purpose of a corpus the remaining steps towards a usable corpus (decision of size, priorities for selection, text picking, ..) are easily taken. Nevertheless, there are two practical matters a corpus builder has to deal with: electronic form and copyright protection of texts.

Sinclair's book was published in 1991, so one could say that the first factor can be ignored due to the rapid pace of technological change. But this can only be applied to certain areas of language which are electronically stored. Generally these are written texts, where a high percentage of all published material already is in machine readable form (books, journals, written conversation, newspapers, …) available through CD-Roms or even online. This development would reduce Sinclair's methods of text input (1991:14) to the following: a) adaption of material already in electronic form and b) conversion by keyboarding. Method b) refers to texts only available in handwritten form (e.g. personal letters) and finally a very important variety of language – spoken text, which has to be transcribed and keyed in. Soon it may be possible to have the computer do the transcription of spoken texts but today this is a major problem corpus builders have to face. This may be best seen in the following figures taken from the BNC which aims at producing a carefully balanced sample of data of British English. Here the proportion of written to spoken text (90 million to 10 million) clearly shows the difficulties of collecting this variety of texts. (Rundell) The second and still unsolved problem is the matter of copyright protection of texts, a sensitive law matter that can slow down the process of corpus creation enormously.

Resulting from these handicaps corpus developers are trapped between two confronting positions: the "opportunistic" approach where all data on can get easily hold of is collected or the "principled" approach where texts are collected following carefully designed specifications (e.g. Brown, LOB). As a way out of this dilemma most "representative" corpora are built by a combination of both methods.

1.4. STEPS IN CORPUS CREATION

After having discussed the practical problems I want to deal with the criteria for the selection of texts. Several factors have to be kept in mind and decisions have to be made: a) written or spoken language, b) formal or literary written texts, c) typicality, d) sample size and overall size. In the following paragraphs these aspects will be discussed in detail concerning the creation of a "representative" corpus.

a) As can be seen above the decision between spoken and / or written language is the most far reaching. Linguists do not agree on the fact that spoken language is a better guide to the fundamental organisation of language, but in order to build a representative corpus, spoken language has to be included, regardless of the different opinions or difficulties explained above. As it is often difficult to gather samples of spoken language corpus builders may be tempted to include "quasi-speech" (film scripts, drama texts, records of public meetings, court cases, …). But these texts only have a limited value, since they are no natural speech acts and can in no way represent natural conversation.

b) The next step is to decide on the range of material of written language. Formal and literary texts (newspaper articles, novels, …) will be more easily collected than informal and ordinary texts (letters, leavelets, …). The main focus should be on the latter since these texts are widespread and typical of everyday prose. These ideal guidelines, however, can hardly be followed and statistic data from the Birmingham Bank of English, which has a clear commitment to a diversity of text-types, shows that about 70% of the texts are from journalistic sources. (Rundell)

c) The most common use of a corpus is to show what is typical and essential for a certain language: it is not only used to compile data for grammars or dictionaries but it can also be a basis for studying literary work. The aim of a corpus builder should be to keep the level of the corpus as general as possible. This would not only mean reduceing the proportion of literary texts which tend to influence everyday language but also leaveing out texts of well-known journalistic writers who tend to have unusual ways of writing. Sinclair says: "This is a minefield of prejudice and misunderstanding. If we are to approach a realistic view of the way in which language is used, we must record the usage of the mass of ordinary writers […]" (1991:17)

So the general guideline for creating a representative corpus is not to sample material from different specialist areas (technical, dialectal, juvenile, …) but to use broadly homogeneous material which is gathered from a variety of sources so that the individuality of a source is obscured in the overall mass of texts. These aspects discussed above deal with the contents of a corpus.

The aspects to be discussed further on will refer to the size of a corpus. The question of size and of finite or infinite form is of prime concern to most researchers. In the past it was a mere ques-

tion of technological possibilities but today the discussion about size is led by the arguments of qualitative and quantitative data drawn from a corpus.

1.5. The Characteristics of Texts in a Corpus

In this chapter I want to summarize the guidelines (Sinclair, 1991:21) corpus developers have to follow when working on texts that are to be collected in the corpus. Corpus builders have to apply certain standards to their texts in order to make it easier for users to work with different corpora and to be able to do the same research on each.

The characteristics of each text should be summarized as follows: it has to be specified if a text is fictional or non-fictional; taken from a book, journal, …, or newspaper; written in formal or informal style; and what is more the author's age, gender and origin should be given. With this information included, the results of any research could be evaluated more precisely. Last but not least, to obey copy-right laws it is equally important to add the full bibliographic information of each text.

Here it is important to note that this information is to be stored separately from the actual text. This procedure will allow different ways of analysis, providing an "equal starting-point" for each researcher. One main reason for this is that the term 'word' is not clearly defined and in no way standardised which leads to various kinds of investigations dealing with a similar field of analysis.

In the early 1990s the Text Encoding Initiative (TEI) started to work on these problems and set rules for the encoding of electronic texts. Their aim was to provide a widely accepted set of encoding standards for corpus-based work:

> "The overall goal is the identification of a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and linguistic information) as well as general architecture (so as to be maximally suited for use in a text database). It also provides encoding conventions for more extensive encoding and for linguistic annotation." (Corpus Encoding Standard, CES1-0.html)

The most recent publication (1996) of the Corpus Encoding Standard (CES) provides the following:

- a set of meta-language level recommendations (particular profile of SGML use, character sets, etc.)
- tag-sets and recommendations for documentation of encoded data;
- tag-sets and recommendations for encoding primary data, including written texts across all genres, for the purposes of corpus-based work in language engineering.
- tag-sets and recommendations for encoding linguistic annotation commonly associated with texts in language engineering, currently including: (segmentation of the text into sentences and words [tokens], morpho-syntactic tagging, parallel text alignment)

(Corpus Encoding Standard, CES1-0.html)

It is important to note that the TEI is an ongoing project and therefore there are areas not dealt with in the CES. Over recent years a large number of additional features have been added to the initial guidelines mostly following the technological development in the area of computer assisted language analysis.

1.6. SIZE AND "STATE OF THE ART" IN CORPUS CREATION

Sinclair, head of the COBUILD team – constructors of a monitor corpus, opts for infinite size: his main argument is the bigger the corpus the more occurrences of a word can be found and the more easily the behaviour of a word in a text can be described. This argument is even of more importance if we begin to classify occurrences in terms of 'uses' or 'meanings', where one 'use' of a word will be more common than an other. (cf. 1991:18) Sinclair also advises corpus creators to build their corpus with whole text samples in order to make it open to a wider range of linguistic research. If there is a need for a smaller corpus or a corpus built out of text samples (as the Brown or the LOB) the large one can always be reduced in size. Another disadvantage of sample corpora is that it may seem as a collection of fragments where only small-scale patterns are available, whereas in a whole text corpus the special patterns of texts are not destroyed.

What Sinclair is referring to is the qualitative analysis of a corpus: the steady adding of new texts allows researchers to look for new words and the shift in meaning of old words. The advantages of these corpora are that they do not give a single "snapshot" of a language but constantly reflect the "state-of-the-art" of a language. The disadvantage, on the other hand, is that these corpora cannot be used for quantitative analysis of language, since they constantly shift in size and are less thoroughly sampled than finite corpora. (McEnery and Wilson, 2fininte.htm)

Current developments in corpus creation show two main branches of interest: huge corpora for lexicographic research and limited and corpora for specialised studies. The best known corpora of the first group is certainly the Bank of English. This corpus is continually under construction, with old texts removed and even greater volumes of new text added. Longman, at the same time, is extending its "Corpus Network" and has added large written and conversational corpora of American English to complement its BNC holdings. Rundell states: "The big lexicographic corpora, then, are heading relentlessly towards a new 1000-million-word benchmark."

The development in the second branch with the introduction of countless new and specialised corpora seems to be even more important and proves a diversification in corpus building. Some of these "newcomers" are for example a corpus of conversational English being collected at the University of Nottingham, and highly specialised corpora. In recent years there has also been an increase in the number of "learner corpora". This development has been encouraged by the success of the first corpus of this kind, the Longman Learner Corpus (sampled in the late 1980s). Recently created corpora are the International Corpus of Learner English (ICLE), a very large corpus of English written by Cantonese learners at the Hong Kong University of Science and Technology, and an exam-based learner corpus at Cambridge University Press. (Rundell).

Most interesting, though, is the increase of corpora in languages other than English since the majority of corpora developed in the past decades were dedicated to the study of English. Over the recent years there have been built corpora for most of the European languages: the most important is a multilingual corpus of 93 million words that has been sampled by the European Corpus Initiative (ECI). On Michael Barlow's corpus web-site links and information to and about corpora in 12 different European languages besides English can be found. (see Reference) The present situation is that all over the world a great deal of work is being done to assemble corpora of countless languages, such as Japanese, Mandarin, or Malay.

2. DESCRIPTION OF CORPORA USED IN THIS PAPER

In this chapter I want to give a short description of the different corpora I am going to use for my work. The main information will focus on which kind of corpus is used and why I settled on using it; the size of the corpus (token /type); the sampling period and the language variety of the texts.

2.1. *ALICE IN WONDERLAND* / *THROUGH THE LOOKING GLASS* BY LEWIS CARROLL (ALICE CORPUS)

This corpus is the basis of my work. It could be labelled as a sample corpus of most limited kind because it is a fictional text written by a single author.

Dialect:	British English
Language variety:	Written
Sampling Period (= date of publication):	1865 / 1872
Size:	Tokens: 56.272
	Types: 3.912

2.2. THE GUARDIAN 1990 – 1993 (GUARDIAN CORPUS)

This corpus again can be seen as a sample corpus representing a special variety (newspaper) of written British English. With the help of these texts I am going to show the influence of Lewis Carroll's books on present day British English.

Dialect:	British English
Language variety:	Written
Sampling Period:	01.01.1990 – 31.12.1993
Size: (approximately)	Tokens: 102.454.976
	Types: 380.813

2.3. TIME MAGAZINE 1989 – 1993 (TME CORPUS)

This corpus will be used in the same way as the Guardian corpus. The difference lies in the dialectal region: this corpus represents present day American English.

Dialect:	American English
Language variety:	Written
Sampling Period:	01.01.1989 – 31.12.1993
Size: (approximately)	Tokens: 10.413.901 Types: 112.581

2.4. LOB CORPUS / FLOB CORPUS; BROWN CORPUS / FROWN CORPUS

These corpora are typical sample corpora of written British and American English. By comparing the wordlists of these corpora to a key-word list derived from the Alice Corpus I am going to show why they are not suitable for the kind of analysis I am going to do.

LOB:

Dialect:	British English
Language variety:	Written
Sampling Period:	1961
Size: (approximately)	Tokens: 1.219.492 Types: 39.652

BROWN:

Dialect:	American English
Language variety:	Written
Sampling Period:	1961
Size: (approximately)	Tokens: 1.197.876 Types: 42.576

FLOB:

Dialect:	British English
Language variety:	Written
Sampling Period:	1991
Size: (approximately)	Tokens: 1.237.428 Types: 45.086

FROWN:

Dialect:	American English
Language variety:	Written
Sampling Period:	1991
Size: (approximately)	Tokens: 1.241.895 Types: 45.446

2.5. BRITISH NATIONAL CORPUS

The BNC is a sample corpus, too. The main differences to the corpora above are its size of about 100 million words and the rather extended sampling period. Its creators intended to give a detailed picture of the English language of recent decades. By comparing the Alice wordlist to the BNC-

written wordlist I tried to specify key-words which may show Carroll's influence on the English language of today.

Dialect:	British English
Language variety:	Written
Sampling Period:	1975 – 1993 (2596 out of 3309 texts)
Size: (approximately)	Tokens: 90.748.880 Types: 377.384

3. CORPUS ANALYSIS

3.1. KEYWORD LIST

This part of the paper summarizes the actual working process. My intention is to outline the influence of Lewis Carroll's books *Alice in Wonderland* and *Through the Looking Glass* on the lexicon of the English language of today. I want to find out which words, phrases or even clauses have been adapted into a general vocabulary. Along to this I want to analyse the collocation and meaning of such words and phrases. All of the work is done with the help of Mike Scott's *Word Smith Tools*, Version 3.00.00.

The first step of my work is to extract key-words from the Alice wordlist. This is done by comparing the Alice wordlist to the BNC-written wordlist. My aim is to produce as much matches as possible, meaning that I adjust the p-value in the keyword settings of *Word Smith Tools* to 0,1 which will produce a maximum result, being 1% in danger of producing a wrong relationship. This is a fairly high standard as in social science a 5% risk is considered acceptable.

The next step to be taken is to select content words from this list, which contains about 1500 words (articles, prepositions, adjectives, verbs, …). This actually is the crucial part of my work because the selection has to be done manually. I decided on the following criteria for the selection of the final keyword list: proper names (Alice excluded); content words (nouns and adjectives) that are used quite often in the Alice text but are not that frequent in the BNC wordlist; and finally nonsense words created by Carroll that also occur in the BNC corpus. Other factors that undoubtedly influenced my decisions are general knowledge, reference literature to the Alice books, and finally curiosity about how certain words (mostly nonsense words) are used today.

The final list contains 98 words of which I want to find out how they are used in present day British and American English. The keyword list is arranged in alphabetical order.

N	Word	6	BOUGH	12	COMFITS
1	ATHELING	7	BRILLIG	13	CONTRARIWISE
2	BANDERSNATCH	8	BROILING	14	CORKSCREWS
3	BARROWFUL	9	BROOCH	15	CURTSEYED
4	BEHEAD	10	CALLAY	16	DAINTIES
5	BOROGOVES	11	CHESSMEN	17	DITTO

18	DODO	45	JABBERWOCKY	72	STIGAND
19	DORMOUSE	46	JURYMEN	73	STOOP
20	DUM	47	LITHE	74	STUPIDER
21	DUMPTY	48	LORY	75	STUPIDEST
22	EAGLET	49	LULLABY	76	SUETY
23	EHT	50	MACASSAR	77	THIMBLE
24	ETTER	51	MALLETS	78	THOUGHTFULLY
25	FERRETS	52	MIMSY	79	TOVES
26	FLAPPERS	53	MOCK	80	TREADING
27	FOOTMAN	54	MORCAR	81	TUMTUM
28	FRABJOUS	55	MUCHNESS	82	TUNE'S
29	FRUMIOUS	56	NOHOW	83	TUREEN
30	GALUMPHING	57	OOP	84	TURTLE
31	GIMBLE	58	OOTIFUL	85	TWEEDLE
32	GNAT	59	OUTGRABE	86	TWEEDLEDEE
33	GRYPHON	60	PORPOISE	87	TWEEDLEDUM
34	GYRE	61	PORTMANTEAU	88	TWIDDLE
35	HADDOCKS	62	PRATTLED	89	TWIG
36	HAIGHA	63	QUADRILLE	90	TWINKLE
37	HARE	64	QUEEREST	91	UGLIFICATION
38	HATTA	65	RATHS	92	UNCLASPING
39	HATTER	66	RILL	93	UNICORN
40	HEARTH	67	SLITHY	94	VORPAL
41	HOOKAH	68	SLUGGARD	95	WABE
42	HUMPTY	69	SNAPPISHLY	96	WALRUS
43	IMPENETRABILITY	70	SNOWDROP	97	WHIFFLING
44	JABBERWOCK	71	SOO	98	WORSTED

3.2. WORDLIST COMPARISON

By comparing the keyword list to the wordlists of different British and American corpora (LOB, FLOB, Brown, Frown) I want to show why these corpora are not suitable for more detailed research and why I have to use newspaper corpora to get a higher number of results for my analysis. The results of this comparison are given in alphabetical order.

KEYWORDS VS. LOB:

1	BROOCH	5	HEARTH	9	TWINKLE
2	DODO	6	MOCK	10	UNICORN
3	FOOTMAN	7	THOUGHTFULL	11	WALRUS
4	HARE	8	TURTLE		

KEYWORDS VS. BROWN:

1	DUMPTY	6	HUMPTY	11	TURTLE
2	ETTER	7	MOCK	12	TWINKLE
3	FOOTMAN	8	QUADRILLE	13	WALRUS
4	HARE	9	THIMBLE	14	WORSTED
5	HEARTH	10	THOUGHTFULLY		

KEYWORDS VS. FLOB:

1	DITTO	5	HARE	9	THOUGHTFULLY
2	DODO	6	HEARTH	10	TUREEN
3	DORMOUSE	7	MOCK	11	TURTLE
4	FOOTMAN	8	THIMBLE	12	TWINKLE

KEYWORDS VS. FROWN:

1	DITTO	5	HEARTH	9	TWINKLE
2	FOOTMAN	6	MOCK	10	UNICORN
3	GNAT	7	THOUGHTFULLY	11	WALRUS
4	HARE	8	TURTLE	12	WORSTED

The results above indicate that Lewis Carroll's books did not have much influence on the English language as it is represented in these sample corpora. What is even more astonishing is that the well known and widely used phrase of the "Mad Hatter's tea party" is not included in any of the corpora. The reason for this minute result can be found in the sampling criteria and the relatively small size of the corpora.

A comparison of the Alice keyword list to the wordlists of *The Guardian* and *Time* will prove that there has been an influence. The lists are sorted by the frequency of the words occurring in the Alice corpus.

KEYWORDS VS. GUARDIAN CORPUS:

N	**Word**	**Freq.**	30	WABE	5
1	TURTLE	56	31	WORSTED	5
2	HATTER	55	32	BRILLIG	4
3	MOCK	55	33	BROOCH	4
4	GRYPHON	54	34	JURYMEN	4
5	DUMPTY	53	35	MIMSY	4
6	HUMPTY	53	36	PORPOISE	4
7	DORMOUSE	39	37	QUADRILLE	4
8	TWEEDLEDUM	33	38	RATHS	4
9	HARE	29	39	THIMBLE	4
10	TWEEDLEDEE	26	40	BANDERSNATCH	3
11	UNICORN	22	41	CHESSMEN	3
12	GNAT	18	42	DITTO	3
13	WALRUS	14	43	JABBERWOCK	3
14	DODO	13	44	MUCHNESS	3
15	FOOTMAN	13	45	TUREEN	3
16	THOUGHTFULLY	11	46	BARROWFUL	2
17	HATTA	10	47	BOUGH	2
18	TWINKLE	8	48	CURTSEYED	2
19	LORY	7	49	DUM	2
20	OOP	7	50	FERRETS	2
21	SOO	7	51	IMPENETRABILIT+	2
22	CONTRARIWISE	5	52	JABBERWOCKY	2
23	GIMBLE	5	53	LITHE	2
24	GYRE	5	54	LULLABY	2
25	HEARTH	5	55	PORTMANTEAU	2
26	HOOKAH	5	56	SNOWDROP	2
27	NOHOW	5	57	TREADING	2
28	SLITHY	5	58	TWEEDLE	2
29	TOVES	5	59	TWIG	2

KEYWORDS VS. TIME CORPUS:

N	**Word**	**Freq.**	6	HUMPTY	53
1	TURTLE	56	7	TWEEDLEDUM	33
2	HATTER	55	8	HARE	29
3	MOCK	55	9	TWEEDLEDEE	26
4	GRYPHON	54	10	UNICORN	22
5	DUMPTY	53	11	GNAT	18

12	WALRUS	14	26	QUADRILLE	4
13	DODO	13	27	THIMBLE	4
14	FOOTMAN	13	28	CHESSMEN	3
15	THOUGHTFULLY	11	30	MUCHNESS	3
16	TWINKLE	8	31	TUREEN	3
17	LORY	7	32	BOUGH	2
19	CONTRARIWISE	5	33	DUM	2
20	GYRE	5	34	FERRETS	2
21	HEARTH	5	35	LITHE	2
22	HOOKAH	5	36	LULLABY	2
23	NOHOW	5	37	STOOP	2
24	WORSTED	5	38	TREADING	2
25	BROOCH	4	39	TWIG	2

The results above clearly present a better picture of the actual situation. The next and final step of my analysis will be to extract kwic-lines from the newspaper corpora to study the usage of some of the words and to prove that they are related to the Alice text.

3.3. Concordance Analysis

Having limited the keywords for each newspaper corpus I want to analyse the usage of some these words and the way in which they are related to Carroll's books. Furthermore, I am going to verify the source of the words and phrases. This will allow me to show if their use was introduced by Carroll.

The method of this analysis will be as follows: firstly I am going to outline the usage and meaning of words and phrases in the Alice text; secondly I will try to provide the background for the words and phrases; thirdly I am going to present the kwic-lines (or at least some of them) from the newspaper corpora; and finally I am going to examine the usage of the keywords in present day English. In this last step I want to analyse how the words have been taken over: by using the phrase of the original text or by adapting a context.

Hatter / Hare / Dormouse:

This is certainly the best known word (today it is a well known idiom) from the Alice text. "Hatter" refers to chapter VII of *Alice in Wonderland* (Gardner 1970: pp. 93): "A Mad Tea Party", where the Hatter, the March Hare and the Dormouse are introduced. They are having a never ending tea party as a punishment for the hatter, who "murdered time" when he sang the song "Twinkle, Twinkle, little Bat!" during a concert given by the Queen of Hearts. In the Alice text the words are used as proper names.

According to the Oxford Dictionary of Idioms [ODI] (1999: 220) the madness of the hatter refers to the fact, that hatters suffered from the effects of mercury poisoning, due to the use of mercury in the production of felt hats. The march hare version of the saying refers to the leaping and running of hares in the breeding season. The dormouse is a tree living rodent, which is named after the Latin word *dormire*, referring to the animal's habit of winter hibernation.

The concordancer produced 166 lines containing “hatter*” from *The Guardian* corpus including 23 versions of the proper name “Hattersley” and 25 versions of a place called “Hatteras”, the remaining 118 lines contained the following versions of the word “hatter”:

Hatter (36,74,77)	proper name
The Mad Hater (27,13)	nick name
hatter/ hatters (5-7)	profession
Hatters (122-124)	Sheffield (Stainless) Hatters (soccer team??)
mad-hatter (32,14)	adjective
Mad Hatter (15-17)	character of the Alice text sometimes + noun
Mad Hatter’s tea party (50-54)	referring to the situation described above, also used with “coffee morning”
mad as a hatter (64-66)	idiom
mad-hattery (166)	noun

```
36     t pressure  groups are unimpressed. Ray Hatter of GLAAS dismisses both the IBA a
74     ning three of the top five were Maurice Hatter, head of  electronics firm IMO P
77     ed to run the CTC  trust, telephoned Mr Hatter. ''Maurice, I've got good news fo
27     ggested me as his  replacement. The Mad Hatter pseudonym came about because my p
13     ually known in the locality as the Mad  Hatter because of his eccentric ideas an
5       man, a mortician, a foundry worker, a  hatter (in the days of steam-and-wooden-
6      ations. Who, for example, wrote: When a hatter will go smatter on  psychology, A
7      ie Buckland, the Monmouth historian and hatter  who's doing a roaring trade in l
121    uted 20 points.  >Earlier the Sheffield Hatters won the women's cup for the four
122    e perimeter.''  The Sheffield Stainless Hatters clinched the women's championshi
123    turns to Nottingham  Forest because the Hatters cannot afford the pounds 300,000
124     as a club offering a warm welcome, the Hatters have become one  of the most dis
32     Kids (Odeon West End, U) a suburban mad-hatter  scientist (Rick Moranis) acciden
14     ly 1993  G2T  PAGE: 12  Beautifully mad hatter  COLIN MCDOWELL  >Obituary: Mr Jo
15     or tea at the Ritz only to find the Mad Hatter doing  the honours, and one has s
16      moment of  weakness, dress up as a Mad Hatter for a computer show in Manchester
17     Ashworth in the  Sporting Life. The Mad Hatter has long been a tiresome apologis
18     f the best prescriptions:  >Did the Mad Hatter have mercury poisoning? This was
50      it seems like something out of  the Mad Hatter's tea party. Certainly it illustr
51     ng run much along the lines of the Mad  Hatter's tea party, but no one could hav
52     ents, like the  participants at the Mad Hatter's tea party with place settings.
53     rever mislaid his invitation to the Mad Hatter's tea party. The highest  complim
54      British Rail again; it is like the Mad Hatter's tea  party.  At Euston station
64     lands, told me  yesterday.  >''Mad as a hatter, and eaten up with rage and anger
65     or Scargill actually. He's as  mad as a hatter, but at least he's . . . oh well,
66     the late Lord Northcliffe who, mad as a hatter, had an  aquarium with piranhas a
166     >Down and out, page 35;  >Gattery, mad hattery, page 36  !9354  SOURCE: The Gua
```

The analysis of the *Time* corpus produced only 13 kwic-lines of which two referred to the concept created in the Alice text. The small number of entries may be due to the small size of the corpus:

```
5      gnats) or make an appearance at the Mad Hatter's tea  party. But by now I've bro
7      ess--beautiful, successful and mad as a hatter, thanks  to the deafening tick of
```

These results lead to the following conclusion: it was not only the word “hatter”, which was adapted to present day English, but a whole concept expressing “insanity” or “madness”. This concept can be expressed in different ways: with a phrase (idiom), a noun, an adjective or as a nick name. (This again shows that the other sample corpora [LOB, FLOB, Brown, Frown] would not have been suitable for this analysis, as the word “hatter” is not found in any of the keyword lists presented above!)

A similar situation can be observed concerning the word “hare”, which in some cases is cited along with the “Mad Hatter” but is also used as an independent phrase, expressing a similar concept, at least in British English. The following lines are taken from *The Guardian*:

```
290    erved to sit beside Alice and the March Hare at that magical table.  During the
291    ; one with the Mad Hatter and the March Hare garlanded with  weeping Mock Turtle
293    n mean  anything. May I quote the March Hare in this connection? Say what you  m
294    ed to view such a move more as a  March hare - mad.  But that is not to say Mr L
296    t. Obviously. The Hatter  and the March Hare torture Dormouse by forcing him int
297    in, in which Akabusi's bounding,  March-hare ecstasy ``I just went loopy'' will
```

The computing of the *Time* corpus, on the other hand, did not produce any suitable results. This could hint at the fact that the “March Hare” concept was not adapted along with the “mad hatter” concept into American English.

Even less used, if it is used at all, is the word “dormouse” and the concept linked to it by Carroll. The only reference I could find was the following line (*The Guardian*):

```
28      The Hatter and the March Hare torture Dormouse by forcing him into a teapot.
```

And even here the protagonists are the Hatter and the Hare. The American English newspaper corpus did not contain the word at all. The reason for this may be that the dormouse is an animal that is only found in Britain (Gardner, 1970:94).

HUMPTY DUMPTY:

The egg-shapred character of Humpty Dumpty as introduced by Carroll originates from an nursery rhyme: “Humpty Dumpty sat on a wall / Humpty Dumpty had a great fall / All the king’s horses and all the king’s men / Couldn’t put humpty together again” (Briggs, 1973:110) Carroll, though, imposes a lot of negative characteristics, such as arrogance or vanity on Humpty Dumpty. The following analysis will show if these characteristics had any influence on the Humpty Dumpty known from the nursery rhyme, and if the name is used along with this context in the English language today. *The Guardian* concordance produced the following lines:

```
19     he Archdeacon of York of  speaking like Humpty Dumpty talking to Alice in his se
20     frost, meteorologists are a little like Humpty Dumpty  in Through the Looking Gl
40     Words mean what I choose them to, says  Humpty Dumpty in Through The Looking Gla
46     ust be a benevolent provider. It is the Humpty Dumpty school of  semantic philoso
48     ishop of Canterbury once likened him to Humpty Dumpty, and the  Archbishop of Yo
54     comprehensible to outsiders.    >As with Humpty Dumpty (this really is Wonderland
```

Time:

```
6       Joycean. They are  more reminiscent of Humpty Dumpty, to whom a word meant what
```

Each of these lines point out that it is the character of Humpty Dumpty as introduced in Through the Looking Glass, either by citing the title itself or by citing his famous declaration “When I use a word it means just what I choose it to mean – neither more nor less.” (Gardner, 1970:269) And it is exactly this concept, which has been taken over along with Carroll’s Humpty Dumpty.

About 90% of the kwic-lines taken from both corpora carry the context provided by the nursery rhyme (it could be rephrased as “some higher force sent out to rescue somebody”). Due to the longer and better tradition of this text, the result is not surprising at all.

TWEEDLEDUM AND TWEEDLEDEE:

These characters also originate from a nursery rhyme: “Tweedledum and Tweedledee / Agreed to have a battle / For Tweedledum said Tweedledee / Had spoiled his nice new rattle. / Just then flew by a monstrous crow / As black as a tar-barrel / Which frightened both the heroes so / That they quite forgot their quarrel.” (Briggs, 1973: 66) Here again Carroll added extra characteristics to the original characters: he gave both a philosophical touch by letting them discuss the Red King’s Dream with Alice. (Gardner, 1970: 238) The concordances will show if this concept has been adapted into the English language. The first line is taken from *Time* and the following four lines from *The Guardian*:

```
1      they are also a bit  like Tweedledee and Tweedledum to many Britons. Heseltine i
6      dings. Local MPs  perform their popular Tweedledum and Tweedledee routines. Spor
7      guests.  >Lords Archer and Healey, like Tweedledum and Tweedledee, meandered out
8      ng. Germany has PR and has  avoided the Tweedledum and Tweedledee swing in polic
9      4 breakfast show.  >The ITC, faced with Tweedledum and Tweedledee brandishing th
```

The lines show that the context added by Carroll has not cought on; it is still the nursery rhyme “two-quarrelling-little-boys” meaning that is conveyed by journalists when citing the names of the two characters.

MUCHNESS:

The words and contexts above refer either to common sayings or to nursery rhymes to which Carroll has added extra meaning. The words or phrases that are discussed in the following paragraphs were coined by Carroll. The analysis of their use in present day English will be most interesting because most of the words are not even mentioned in standard (students’) dictionaries. Some of the words, however, may go back a long way in time – as in the case of the word “muchness” – but I tend to believe that the influence of Carroll is far greater than that of the original etymology.

The idiom I refer to is “much of a muchness” meaning “very similar, nearly the same”. “Muchness” is a Middle English word (meaning “large size, bigness”) and is hardly used outside this phrase in today’s language. (ODI, 1999: 236) As I have not found any other reference to this phrase I claim that it was coined by Carroll. This is the passage of *Alice in Wonderland* where the phrase is introduced:

> “The Dormouse had closed its eyes by this time, and was going off into a doze; but, on being pinched by the Hatter, it woke up again with a little shriek, and went on: `--that begins with an M, such as mouse-traps, and the moon, and memory, and muchness-- you know you say things are "much of a muchness"--did you ever see such a thing as a drawing of a muchness?” (Gardner, 1970: 103)

The following kwic-lines show how the word (phrase) is used in *The Guardian* (lines 1-9) and in *Time* (lines 10-11):

```
1     eep frying and found them all much of a muchness  and perfectly adequate for the
2      players on the  European tour are of a muchness on the practice tee. Ballestero
3      fifth from bottom shows  how much of a muchness the division is. Though the har
5     think of the Baltic tongues as much of a muchness. But there is very little con
6      the same conditions they are much of a muchness.  >Galicia finished at three mi
7     ally so clear-cut? Both  look much of a muchness, part of the great Middle Engla
8           >THE regulars saw it as much of a muchness: ``We were here before it was
9     oint that these machines are much of a  muchness. ''You know, a motherboard from
10    for the eyes--a delirious too much of a muchness.               To Rodriguez, t
11                                      A Muchness Of Maleness
```

From these lines one can see that the concept invented by Carroll has been take on into the English Language of today. Even Gardner mentions this fact in his annotations to the Alice text. (1970:103)

SLITHY:

This is a nonsense word taken from the Jabberwocky poem in *Through the Looking Glass*. There are two different interpretations for this word. One is by Humpty Dumpty: "…'slithy' means 'lithe and slimy'. 'Lithe' is same as 'active'." (Gardner, 1970: 271) The Oxford English Dictionary lists "slithy" as a variant of "sleathy", an obsolete word meaning slovenly. (Gardner, 1970: 194) But the concordance taken from *The Guardian* seems to convey the first meaning:

```
1     First Tuesday  documentaries in which slithy coves with lawyer written all over
```

CURIOUSER AND CURIOUSER:

This comparative is coined by Alice when she starts to grow after eating a cake (see Gardner, 1970: 35). The author even makes fun of Alice not being able to remember the correct grammar. Today the word is used in the phrase "curiouser and curiouser" which is the actual phrase Alice uses in the text. The meaning of this phrase is "something is surprising". The following kwic-lines are taken from *The Guardian*:

```
5     real life they were snorting coke? And, curiouser and curiouser, why were  the f
10    the Pope occasionally makes mistakes.  >Curiouser and curiouser, up from the Lab
19     were snorting coke? And, curiouser and curiouser, why were  the fans always eno
20    >THE championship becomes curiouser and curiouser, writes Neil Robinson. In  som
21    the pope occasionally makes mistakes.  >Curiouser and curiouser, up from the Labo
22    much for Cameroon at 38; curiouser  and curiouser.  The emergence of Cameroon an
```

A concordance of the word in the *Time* corpus did not produce any results.

BEST BUTTER:

The phrase is used by the Hatter when complaining about a hint by the March Hare on improving the works of his clock: " `Two days wrong!' sighed the Hatter. `I told you butter wouldn't suit the works!' he added looking angrily at the March Hare. `It was the *best* butter,' the March Hare meekly replied." (Gardner, 1970: 96) Today this phrase is used to point out that something or

somebody is very special in a certain aspect. The kwic-lines are taken from *The Guardian* (The Time corpus did not contain this phrase):

```
1     bing  scarecrow.  The cast are the best butter - Dorothy Tutin, Helen Cherry, Ro
2     s. It is lusciously cast. Only the best butter as the Mad Hatter,  who would hav
```

OUTRODUCTION

The analysis above indicates that the influence of Carroll's books is far greater on British English than that on American English. There are several explanations: Carroll's books were published quite late in this century in America (*Alice in Wonderland*: 1971 and *Trough the Looking Glass*: 1963), so that they could not influence people and language as long as in Britain. Another, more logical, explanation for the analytical results, though, may be the size of the *Time* corpus, being just one tenth of the size of the *Guardian* corpus. Another factor to be taken into consideration is that *The Guardian* and the *Time Magazine* cannot be regarded as an "equal" source, with the first being a daily newspaper and the second a weekly magazine.

Nevertheless, the corpus analysis has proved that Lewis Carroll's books *Alice in Wonderland* and *Through the Looking Glass* do have a considerable influence on present day American and British English. This paper, however, only scratches the surface of the topic and I will try to provide more detailed results in my talk in June.

REFERENCE

BOOKS:

- Briggs, Raymond (1973), The Mother Goose Treasury, Harmondsworth: Picture Puffin.
- Gardner, Martin, ed. (1970), The Annotated Alice. Lewis Carroll, Revised Edition, Harmondsworth: Penguin.
- Sinclair, John (1991), Corpus, Concordance, Collocation. Oxford: OUP.
- Speake, Jennifer, ed. (1999), The Oxford Dictionary of Idioms, Oxford: OUP.

CORPORA:

- Alice Corpus: Alice in Wonderland & Through the Looking Glass by Lewis Carroll, PROJECT GUTENBERG AND DUNCAN RESEARCH SHAREWARE © 1991, Klug Helmut W., ed (1999), Karl-Franzens-University, Graz, Austria.
- BNC Wordlist, Mike Scott's Web, http://www.liv.ac.uk/~ms2928/index.htm
- Brown Corpus: (1963-1964), W. Nelson Francis, Henry Kucera, Brown University, USA.
- FLOB Corpus : Freiburg-LOB Corpus of British English, (1991), Christian Mair, Albert-Ludwigs-Universität Freiburg.
- Frown Corpus : Freiburg-Brown Corpus of American English, (1991), Christian Mair, Albert-Ludwigs-Universität Freiburg.
- Guardian Corpus: The Guardian 1990 – 1993, © 1993. Contact Guardian Newspapers Ltd.
- LOB Corpus: Lancaster-Oslo/Bergen (LOB) Corpus, (1970-1976), Stig Johansson & G.N. Leech, University of Oslo & University of Lancaster.
- Time Corpus: Time Magazine 1989 – 1993, © 1994 Compact Publishing, Inc. © 1994 TIME Magazine Inc. Co.

INTERNET:

- Michael Barlow, Corpus Linguistics: http://www.ruf.rice.edu/~barlow/corpus.html#Doc
- Tony McEnery and Andrew Wilson, Corpus Linguistic:
- http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm
- Michael Rundell, The future of the corpus, and the corpus of the future:
- http://www.ruf.rice.edu/~barlow/futcrp.html
- Corpus Encoding Standard:
  http://www.lpl.univ-aix.fr/projects/multext/CES/CES1.html

SOFTWARE:

- WordSmith Tools, Version 3.00.00, © Mike Scott 10/12/98.