

# Age Tagging and Word Frequency for Learner's Dictionaries

**Hanhong LI & Alex C. Fang**

Dialogue Systems Group  
Department of Chinese, Translation and Linguistics  
City University of Hong Kong

Presented in **American Association for Corpus Linguistics 2009**

# Outline

- § Introduction
- § Methodology and Resources
- § Core Age List from BNC Age Sub-corpora vs Defining Vocabularies in Learner's Dictionaries (Oxford & Longman)
- § Conclusion

# 1. Introduction

§ Word Frequency in Current Learner's Dictionaries

§ Word frequency information has been an indispensable part of English learner's dictionaries. The review of the current major electronic learner's English dictionaries show that word frequency information is mainly based on the raw frequency counting without making full use of tagging for age information in corpora.

# Word Frequency in LDOCE 5th

§ *The Longman Dictionary of Contemporary English 5<sup>th</sup> edition (LDOCE5)* includes Longman Communication 3000 words which are based on statistical analysis of the 390-million-word Longman Corpus Network.

§ They are common English words and present the core of English and show students which words are the most important to learn and study in order to communicate effectively in both speech and writing. (Bullon and Leech 2007:1)

§ Longman Defining Vocabulary (LDV) enjoys a long history since 1978 and starts as a controlled vocabulary and has been regarded as core vocabulary (Quirk 1978, Lee 2001)

§ Words in LDV are constantly being researched and checked to make sure that they are frequent in the Longman Corpus.

§ Longman defining vocabulary (LDV) of around 2000 words has been used to write all the words in LDOCE5.

§ Word Selection Criteria for Longman  
Communication 3000

A) frequency B) range

§ Word Selection Criteria for LDV  
frequency

LDOCE 5

Home | Copy | Print | My Dictionary | Help

# Longman

Dictionary of Contemporary English

Dictionary | Activator | Grammar | Exercises | Vocabulary Trainer | Teacher Resources | Pop up Dictionary | Writing Assistant

wonderful

Spell check  
Phrase search  
Pronunciation search  
Advanced search

A-Z CULTURE

**won-der-ful** **S1** **W2** /wʌndəfəl \$ -də- / adjective

Word family

**1** making you feel very happy **SYN** great:  
☛ *We had a wonderful time in Spain.*


**2** making you admire someone or something very much **SYN** amazing:  
☛ *It's wonderful what doctors can do nowadays.*

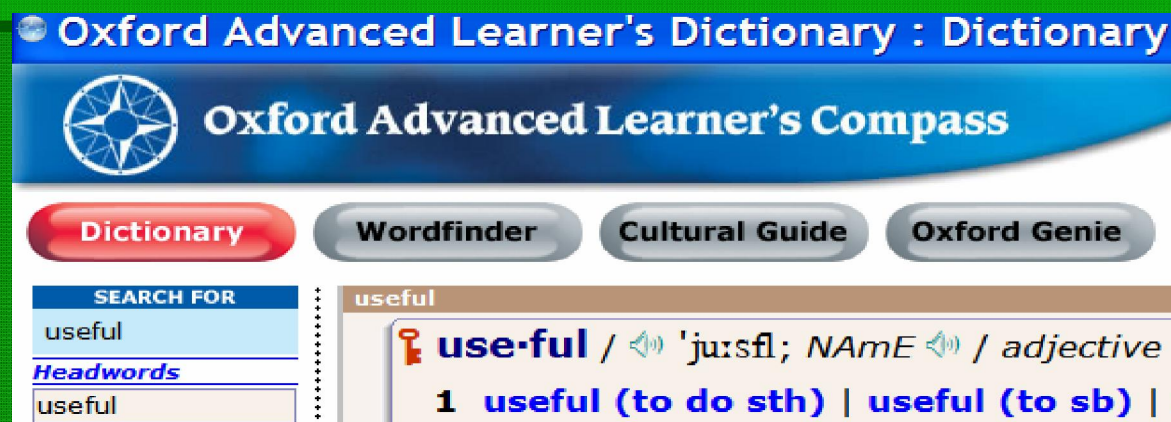
**COLLOCATIONS**  
Collocations from this entry  
Collocations from other entries  
Collocations from the corpus

**THESAURUS**  
Thesaurus  
Longman Language Activator  
Word sets


**PHRASE BANK**

# Word Frequency in OALD7

§ The words of the **Oxford 3000** are shown in the main section of the dictionary with a preceding key symbol . They are the defining vocabulary for *Oxford Advanced Learner's Dictionary 7<sup>th</sup> edition* (OALD7 2005).



Oxford Advanced Learner's Dictionary : Dictionary




 Oxford Advanced Learner's Compass

Dictionary Wordfinder Cultural Guide Oxford Genie

SEARCH FOR  
useful

Headwords  
useful

useful

 **use-ful** /  'ju:sfɪ; NAmE  / adjective

**1** **useful (to do sth) | useful (to sb) |**



## § Word Selection Criteria for Oxford 3000

A) frequent

B) widely distributed

in a wide range of text types

C) familiar

# Word Frequency in MED2

§ 7,500 words printed in red are the **core vocabulary** of English in the *Macmillan English Dictionary 2<sup>nd</sup> edition* (MED2 2007)

All red words have a 'star rating':

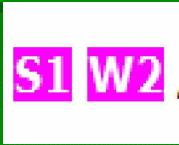



- ★★★ the 2,500 most common and basic English words, such as *easy, go, have, house*
- ★★ very common words, such as *behave, frighten, intelligence, occasional*
- ★ fairly common words, such as *boil, cruelty, farming, metric*

# Word Frequency in CALD3

§ In *Cambridge Advanced Learner's Dictionary 3<sup>rd</sup> edition* (2008), Word frequency is indicated as follows but there is no further detailed explanation of the word selection criteria behind the word frequency symbols.

## ■ Word frequency symbols.

Symbol	Meaning
E	Essential: a common, useful and important word to know.
I	Improver: a word to help you improve beyond basic English.
A	Advanced: a word to make your English sound advanced.

Learner's Dictionary	Frequency Marking	Defining Vocabulary	Criteria
LDOCE5		Yes	frequency
OALD7		Yes	frequency, wide distribution in text types, Familiarity
MED2		Yes	frequency
CALD3		No	?

# Frequency vs Age

- § Frequency has been indispensable for word selecting in defining vocabulary. Is it enough?
- § Early experiments (Carroll & White 1973) have demonstrated that word retrieval in long-term memory is much more influenced by the age of acquisition than word frequency.
- § For English learners in non-English speaking countries, it is necessary for them to know what words are used by native speakers of different ages besides those frequent words.
- § Core vocabulary are not only those words with high frequency but also those words with an even distribution in different age groups. Therefore, the research of age information for word frequency will be of significance for English learning and teaching, lexicography, and core vocabulary research.

# Current Project

- § explore core vocabulary selection by the word frequency distribution in different age groups.
- § compare the core lists selected by both age and frequency parameters with other core lists in learner's dictionaries which were mainly selected by raw frequency.

## 2. Methodology and Resources

### § Corpora:

BNC XML

CHILDS

LUCY

LCCW

POW

### § Core Vocabulary Lists:

defining vocabulary in learner's dictionaries

Longman Defining Vocabulary

Oxford Defining Vocabulary

### § Statistical Method: Carroll's U

## 2.1 Original Age Tagging in BNC XML

```
- <bncDoc xml:id="F7A">
- <teiHeader>
.....
<person ageGroup="Ag4" xml:id="PS1L8"
  role="unspecified" sex="m" soc="UU" dialect="NONE"
  educ="X">
  <age>50+</age>
  <persName>Roger</persName>
</person>
.....
</teiHeader>
```



- <u who="PS1L8">

- <s n="1">

<w c5="UNC" hw="erm" pos="UNC">Erm</w>

<w c5="VM0" hw="can" pos="VERB">can</w>

<w c5="PNP" hw="i" pos="PRON">I</w>

<w c5="AV0" hw="also" pos="ADV">also</w>

...

</u>

# Modification of Age Tagging in BNC XML

§ `<u who="speaker's ID">`

were replaced with a more detailed uniform pattern:

```
< u ageGroup="X" xml:id="X" role="X"  
sex="X" soc="X" dialect="X"  
firstLang="X" educ="X" n="X">
```

## 2.2 Age Group Statistics in Original BNC XML

AgeGroup	Spoken	Written	Total Token
Ag0(0-14)	385,233	59,559	444,792
Ag1(15-24)	594,400	542,578	1,136,978
Ag2(25-34)	1,120,516	2,267,123	3,387,639
Ag3(35-44)	1,075,749	6,726,931	7,802,680
Ag4(45-59)	1,638,364	7,230,715	8,869,079
Ag5(60)+	1,137,433	5,156,077	6,293,510
Total	5,951,696	21,982,983	27,934,679

# Age Group Subcorpora Sampled from BNCXML

Age Groups	Spoken	Written	Token	Type
Ag0 (0-14)	302618	59559	362177	16080
Ag1 (15-24)	500397	500868	1001265	24932
Ag2 (25-34)	510863	513666	1024529	25219
Ag3 (35-44)	525850	501106	1026956	26572
Ag4 (45-59)	501575	506572	1008147	26582
Ag5 (60)+	505803	507623	1013426	24796

# Supplementary Children Spoken Corpora

supplementary corpora	token	age range
Polytechnic of Wales Corpus (POW) (Tim O'Donoghue, Clive Souter 1989 )	60,717	6-12
Child Language Data Exchange System(CHILDES): British English (MacWhinney, B. 2000)	788,324	0-7

# Supplementary Children Written Corpora

supplementary corpora	token	age range
LUCY Children (Geoffrey Sampson 2005)	30,000	9-12
Lancaster Corpus of Children's Project Writing (LCCPW) (Roz Ivanic and Tony McEney 2001)	82,396	8-11

# Final Composition of Agegroup0

Components of Ag0 in Age-Group Corpus	Token
BNC WAg0	59559
LUCY Children Writing Part (LUCY)	25995
Lancaster Corpus of Children's Project Writing (LCCPW)	82396
BNC SAg0 (sampled)	302618
CHILDES(Howe)	15306
CHILDES (Wells)	109411
CHILDES(Manchester –anne-warr-car)	405699
<b>Total</b>	<b>1000984</b>

# A Balanced Age-Group Corpus

Age Groups	Token	Type
Ag0 (0-14)	1001113	16081
Ag1 (15-24)	1001244	24932
Ag2 (25-34)	1024438	25219
Ag3 (35-44)	1026930	26572
Ag4 (45-59)	1007872	26582
Ag5 (60)+	1013412	24796



# Core Vocabulary List

Core List	Type (including derivatives)	Selection Criteria
Longman Defining Vocabulary (LDV5, 2009)	2787	raw frequency
Oxford Defining Vocabulary (ODV7, 2005)	3627	raw frequency, distribution in wide range text types, familiarity

# Compare Core Lists from Age-Group Corpus and Learner's Dictionaries

§ AgeList1 vs LDV

§ AgeList2 vs ODV

## 2.3 Core Age List Computation Method: *frequency + age*

Carroll's  $U_m$

$$U_m = (1,000,000/N)[FD_2 + (1-D_2)f_{min}]$$

where

$F$  = the total frequency of the given word in the corpus

$N$  = the total number of tokens in the corpus

$D_2$  = Carroll's Dispersion Index (see 1.3.2)

$$f_{min} = (\sum S_j f_j) / N$$

$f_j$  = the sub-frequency of a given word-type in category  $j$  ( $j \sim 1, 2, \dots, n$ )

$S_j$  = the number of tokens in category  $j$

# Age Group Database

	hw	frqAg0	frqAg1	frqAg2	frqAg3	frqAg4	frqAg5	range	frqSum	coverage	CC	D	U
1	be	46433	52230	50513	53282	51788	51831	6	306077	5.04	5.04	1.00	50364.07
2	the	24731	44814	46939	49126	50085	53621	6	269316	4.43	9.47	.99	43832.38
3	i	47792	34525	23964	30351	21281	25219	6	183132	3.01	12.49	.98	29566.31
4	and	17796	23979	24977	23443	25854	28699	6	144748	2.38	14.87	.99	23723.03
5	it	26351	21170	21690	21090	21936	20961	6	133198	2.19	17.06	1.00	21885.08
6	to	14273	22688	22678	24251	23578	24438	6	131906	2.17	19.23	.99	21576.73
7	a	21250	20287	21424	22549	21205	22659	6	129374	2.13	21.36	1.00	21290.91
8	you	20676	20416	22740	21385	20666	18863	6	124746	2.05	23.42	1.00	20521.79
9	of	3694	21074	21456	22446	23718	25214	6	117602	1.94	25.35	.95	18505.71
10	have	15572	17458	17602	18723	17387	17269	6	104011	1.71	27.06	1.00	17112.37
11	that	20526	16220	16186	16883	16582	16652	6	103049	1.70	28.76	1.00	16929.56
12	in	12506	14916	15092	16440	16071	17127	6	92152	1.52	30.28	1.00	15138.18
13	not	17100	15566	14645	15633	13788	13981	6	90713	1.49	31.77	1.00	14910.84
14	he	10090	15059	17428	14342	12930	14729	6	84578	1.39	33.16	.99	13844.05
15	do	17609	13775	13683	13454	11699	10943	6	81163	1.34	34.50	.99	13282.67
16	they	7894	10510	11775	12013	12257	13617	6	68066	1.12	35.62	.99	11139.51
17	she	5097	9312	12855	8199	8417	5790	6	49670	.82	36.44	.98	8007.78
18	on	11004	7100	8119	7810	7539	7355	6	48927	.81	37.24	.99	8006.10
19	we	6508	6256	7489	8938	8320	8063	6	45574	.75	37.99	1.00	7476.66
20	get	12177	7503	6439	6753	5710	5818	6	44400	.73	38.72	.98	7166.31
21	there	16199	4891	5770	5556	5289	6386	6	44091	.73	39.45	.93	6846.35
22	yeah	17704	8038	5349	5420	3925	4475	6	44911	.74	40.19	.91	6808.93
23	go	14090	6289	5452	5386	4781	5011	6	41009	.68	40.86	.95	6442.71
24	for	3913	7032	6897	7240	7287	6734	6	39103	.64	41.51	.99	6383.21
25	no	18389	5950	4424	4616	4137	5065	6	42581	.70	42.21	.89	6347.23
26	what	8547	6507	6005	5799	5304	4141	6	36303	.60	42.80	.99	5906.29

## 2.4 Age Lists

Defining Vocabulary	Word Number	Criteria	Cumulative Coverage (%)
Age List1	2787	Frequency, Age Distribution (U)	87.98
Age List2	3627	Frequency, Age Distribution (U)	89.12

### 3. Data Analysis

Defining Vocabulary	Word Number	Criteria	Cumulative Coverage(%)
Longman	2787	Frequency	79.76
Age List1	2787	Frequency, Age (U)	88.06
Oxford	3627	Frequency, text types, familiarity	86.31
Age List2	3627	Frequency, Age (U)	89.41

# 3.1 Coverage in Age Groups: AgeList1 vs Longman Defining Vocabulary

**Group Statistics**

	group	N	Mean	Std. Deviation	Std. Error Mean
coverage	Longman	6	79.7533	2.23738	.91341
	AgeList1	6	88.0550	.75818	.30952

§ The mean coverage of AgeList 1 ( $M = 88.06$ ,  $SD = .76$ ,  $n=6$ ) is larger than the mean coverage of Longman defining vocabulary ( $M = 79.75$ ,  $SD = 2.24$ , ) by 8.31%. The difference is significant with  $t(6.133) = 8.608$ ,  $p = 0.000 < 0.05$ .



## 3.2 Coverage in Age Groups: Age List2 vs Oxford Defining Vocabulary

**Group Statistics**

	group	N	Mean	Std. Deviation	Std. Error Mean
coverage	Oxford	6	86.3033	1.14887	.46902
	AgeList2	6	89.4017	.87305	.35642

§ The mean coverage of AgeList2 (M = 89.40, SD = .87, n=6 ) is larger than the mean coverage of Oxford defining vocabulary (M = 86.30, SD =1.15, ) by 3.10%. The difference is significant with  $t(9.33) = -5.26, p = 0.000 < 0.05$ .

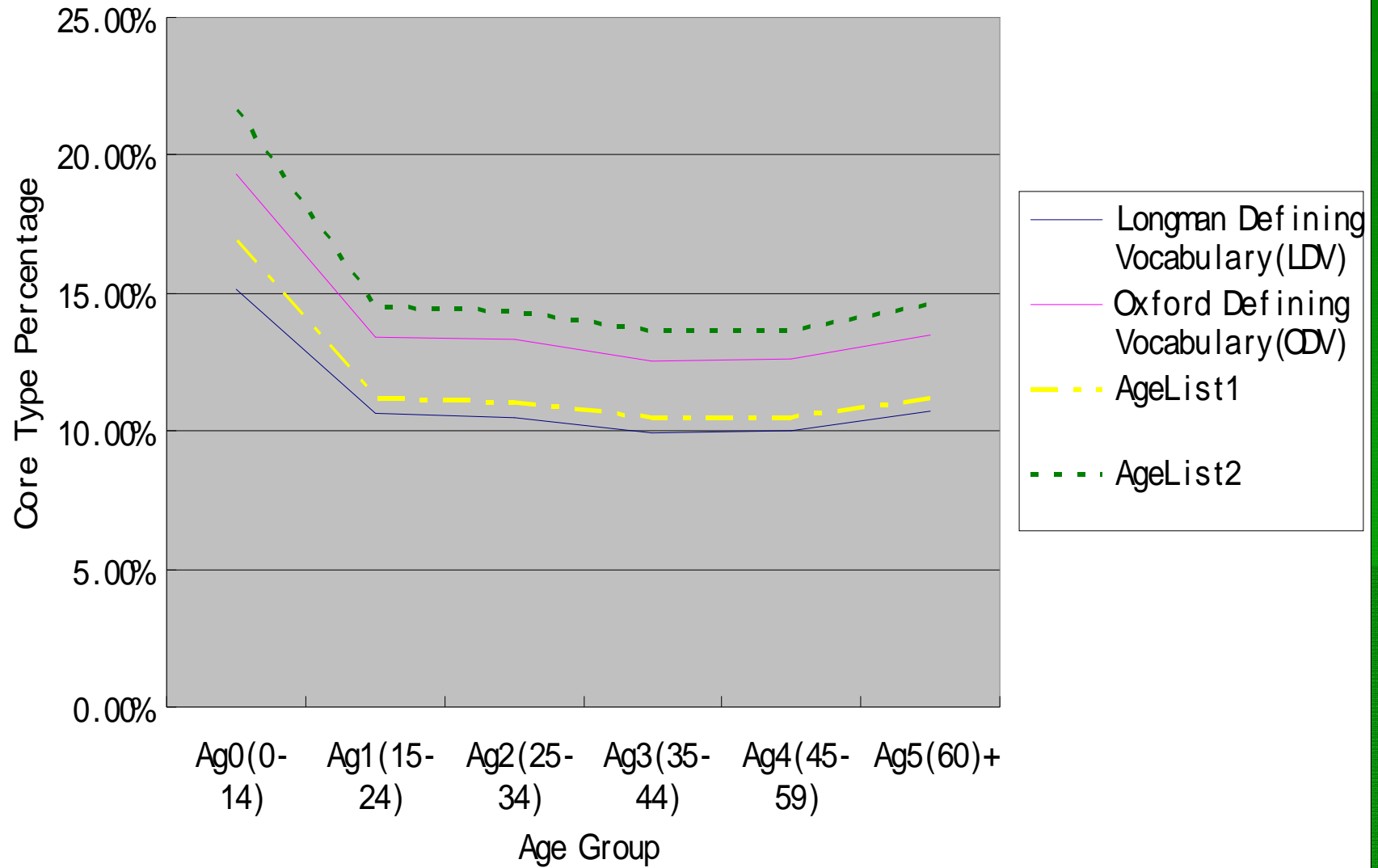
## 3.3 Core Type Percentage

- § Core type percentage = type number of the core vocabulary in an age group/ total word type number in an age group
- § It indicates the proportion of core vocabulary in the total vocabulary size of different age groups of speakers.

# Core Type Percentage Distribution in Age Groups

Core Lists	Ag0(0-14)	Ag1(15-24)	Ag2(25-34)	Ag3(35-44)	Ag4(45-59)	Ag5(60)+
Longman	15.13%	10.63%	10.52%	9.92%	10.02%	10.69%
Oxford	19.35%	13.40%	13.30%	12.55%	12.64%	13.46%
AgeList1	16.85%	11.17%	11.04%	10.48%	10.47%	11.22%
AgeList2	21.57%	14.54%	14.36%	13.64%	13.62%	14.58%

Core Lists in Age Groups



## 4. Conclusion

- § When we select core vocabulary, the combined parameters of word frequency and its distribution in different age groups can achieve higher coverage than frequency only, and even better than the combination of frequency and distribution in text types.
- § Young age groups rely more on core words in their daily communication than adults .
- § People from different age groups tend to use more core words selected on frequency-age basis than those from Longman or Oxford Defining Vocabulary which are selected on frequency or frequency-range basis.

# Future Work

- § It is necessary to test the established Age List in other corpora.
- § More detailed and specific age information in corpora tagging will facilitate our reach in language acquisition and psycholinguistic studies.
- § Frequency marking integrated with native speakers' age information in learner's dictionaries will be helpful for EFL.

# Reference

- § Bullon, S., & Leech, G. (2007). Longman Communication 3000 and the Longman Defining Vocabulary. In *Longman Communication 3000* (pp. 1-7). Harlow: Pearson Education Limited.
- § Carroll J. B., & White, M. N. (1973). Word frequency and age-of-acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, 25, 85-95.
- § Carroll, J. B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index (SFI). *Computer Studies in the Humanities and Verbal Behavior*, 3, 61–65.
- § Carroll, J. B., Davies, P. & Richman, B. (1971). *The American Heritage Word Frequency Book*. Boston: Houghton Mifflin.
- § Lee, D. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics* 29 (3), 250-78.
- § Quirk, R. (1978). Preface to *Longman Dictionary of Contemporary English 1st edition*. London: Longman.



## Dictionaries:

- § *Cambridge Advanced Learner's Dictionary 3<sup>rd</sup> edition* (cd-rom), 2008.
- § *Longman Dictionary of Contemporary English 5th edition* (cd-rom), 2009.
- § *Macmillan English for Advanced Learners 2<sup>nd</sup> edition* (cd-rom), 2007.
- § *Oxford Advanced Learner's Dictionary 7th edition* (cd-rom), 2007.

## Corpora

- § British National Corpus XML edition (BNC)
- § Child Language Data Exchange System (CHILDES):
- § Lancaster Corpus of Children's Project Writing (LCCPW)
- § LUCY Children
- § Polytechnic of Wales Corpus (POW)