HAL open science

But what's wrong with it? Corpus linguistics, helping non-linguists find order in a fuzzy world.

Alex Boulton, Myriam Pereiro

To cite this version:

Alex Boulton, Myriam Pereiro. But what's wrong with it? Corpus linguistics, helping non-linguists find order in a fuzzy world.. M. Pereiro & H. Daniels. Le désordre. Grendel, n° spécial., Nancy: AMAES., pp.161-185, 2008. hal-00327220v2

HAL Id: hal-00327220

https://hal.science/hal-00327220v2

Submitted on 17 May 2009 (v2), last revised 26 Jun 2024 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

But what's wrong with it? Corpus linguistics, helping non-linguists find order in a fuzzy world.

Alex Boulton & Myriam Pereiro. 2008. In In M. Pereiro & H. Daniels (eds.) *Le Désordre. Grendel*, n° special, p. 161-185.

alex.boulton@univ-nancy2.fr

CRAPEL – ATILF/CNRS, Nancy-Université

3 place Godefroi de Bouillon
BP 3397
54015 Nancy – cedex
France

Résumé

Lieu : une classe d'anglais quelque part en France.

ELÈVE: What's the difference between *almost* and *nearly*?
PROFESSEUR: There isn't any.
ELÈVE: So why did you change it in my homework?
PROFESSEUR: ???

En dépit des multiples ressources à leur disposition, les apprenants en langue étrangère ont souvent du mal à trouver une réponse à des questions de ce genre. Nous essayons de démontrer que ni l'intuition des enseignants, ni les dictionnaires de tous ordres ne parviennent à distinguer clairement les différences entre *almost* et *nearly*. Or, une synonymie parfaite est rare, voire inexistante. Comment donc élèves et enseignants pourraient-ils trouver un certain ordre dans le chaos (pour ne pas dire le désordre) que représente une langue ? Les TIC et la linguistique de corpus permettent aujourd'hui d'observer la langue dans son utilisation. Nous prenons *almost* et *nearly* (voir aussi Kjellmer 2003) comme exemples afin de montrer que des enseignants et des apprenants, même des non spécialistes, peuvent avoir recours à de grands corpus libres d'accès pour découvrir par eux-mêmes comment fonctionne une langue.

Cet article s'appuie principalement sur l'interface VIEW du British National Corpus pour comparer la fréquence, la répartition selon les genres de texte et les collocations de *almost* et *nearly*. Il est relativement facile de repérer des différences mais on peut se demander si tout cela serait utile à des non spécialistes. Bien qu'il n'y ait eu pour l'instant qu'un petit nombre d'études empiriques portant sur cette question (voir Boulton 2006, 2007), il y a de bonnes raisons théoriques de penser que l'apprentissage à partir de corpus (*data-driven learning* selon Johns 1991a, 1991b) peut promouvoir l'autonomisation de l'apprenant et accroître ses chances de développer une culture langagière (cf. Gremmo 1995), surtout à un niveau d'études universitaires.

PLEASE NOTE: THIS IS A PRE-PRINT VERSION AND MAY CONTAIN DIFFERENCES WITH THE FINAL PUBLISHED VERSION

Introduction

France's deep-rooted interest in linguistic accuracy is well-known (Bonnet & Levasseur, 2004). Most French teachers of English hope their English is close to that of an educated native speaker and wish to bring their pupils' English as near to it as possible. On the other hand, one survey (Tsui, 2005: 338-339) found that teachers' and learners' most frequent questions were "about lexical items that teachers take to be synonyms or near-synonyms" and where they "have problems explaining to students their difference in meaning or usage."

What follows then stems from an imaginary but highly likely scenario. What would be a teacher's reply to a student who would like to know the difference between *almost* and *nearly*?

1 A traditional approach

When faced with a question about the language they teach, some teachers would perhaps trust their intuition. They are experts, after all, especially if the teacher is a native English speaker. Others might wish to look up both adverbs in a dictionary, relying on the written word. We asked thirty-one trainee teachers to fill in a questionnaire (Appendix 1) which would give us an insight of how helpful intuition can be. These trainees have passed a highly competitive examination which attests of their mastery of the English language and are now teaching two classes under the guidance of mentors and teacher trainers. One of these trainees is English.

We first asked if there were any differences between the two adverbs. Twenty-one trainees, the native speaker included, said there were, ten said not. Those who believed there was a difference gave the following explanations:

The native speaker's intuitions:

- 'Almost' is more formal
- 'Nearly' is more British than 'almost'
- 'Nearly definitely' and 'nearly certainly' sound wrong
- 'Almost' is nearer to the aim than 'nearly'

Non native speakers' intuitions:

- 'Nearly' is closer and more precise
- 'Almost' is more often used with a negative
- The nature of the word that follows will determine the use of one or the other
- The root gives a clue: al-most means we are dealing with a quantity, near-ly, a place or a distance

When faced with a series of closed questions, they became more certain that there were differences between *almost* and *nearly*. Figures 1, 2 and 3 below show that, just as it was with their first intuition, there is a variety of replies.

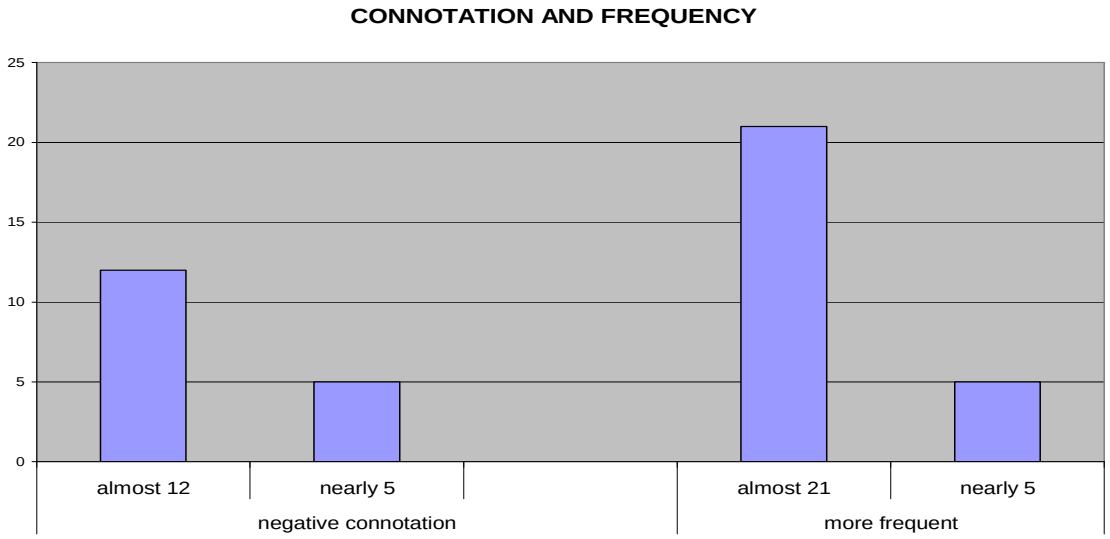


Figure 1. Distribution of answers on whether one adverb is more negatively connoted or more frequently used

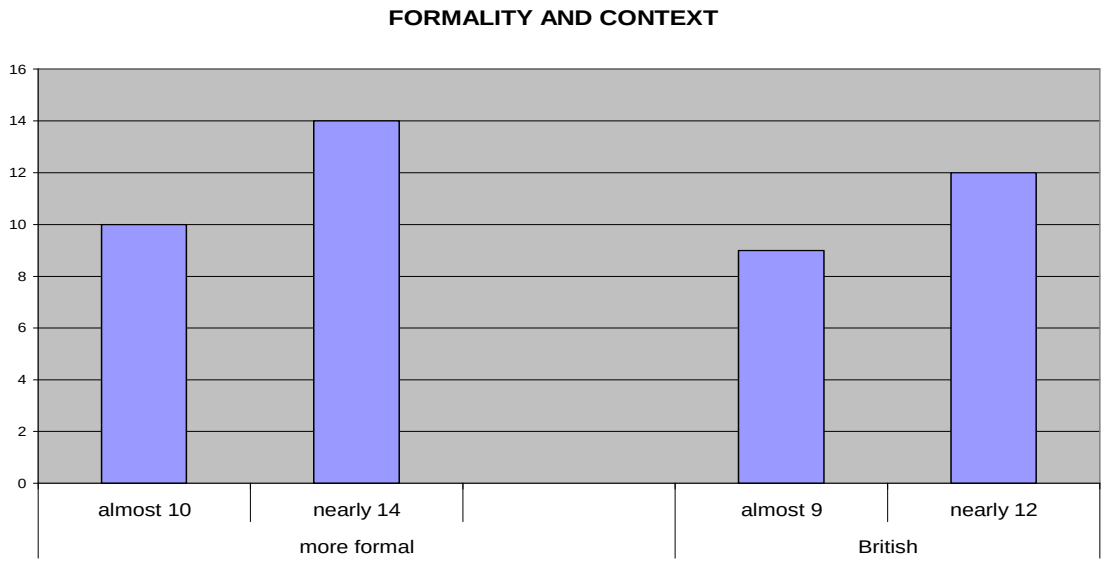


Figure 2. Distribution of answers on whether one adverb is more formal or more typical of one country

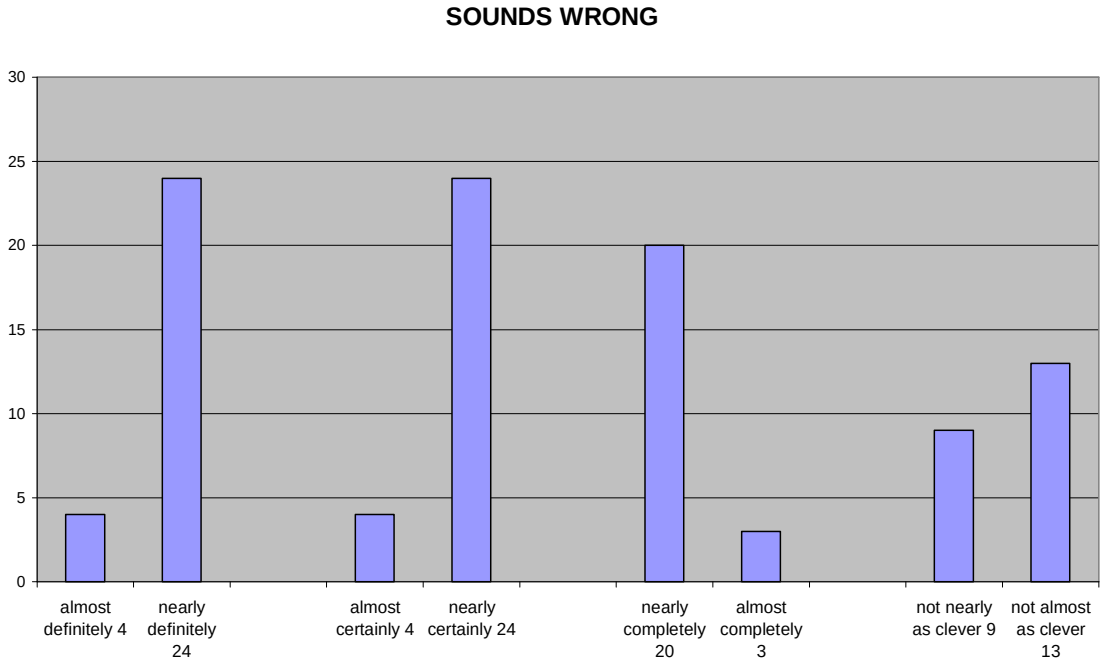


Figure 3. Distribution of answers on combinations that sound wrong

The various replies indicate that teachers do not agree on what sounds wrong, which means that they cannot safely rely on their intuitions. And if they try to, they might present erroneous or relative truths as definite ones, which is clearly unhelpful in the short term and may later result in confusion as learners encounter examples which do not fit the erroneous rule.

However, teachers who prefer to rely on dictionaries are not sure to do better than those who rely on their intuition, as both adverbs are often presented as synonymous.
This is the case with two common dictionaries for native speakers of English:

Merriam-Webster dictionary online

Almost very nearly but not exactly or entirely <we're *almost* there>	Nearly **1:** in a close manner or relationship <*nearly* related> **2 a**: almost but not quite <*nearly* identical> <*nearly* a year later> **b :** to the least extent <not *nearly* as good as we expected>

The Shorter Oxford English Dictionary (1983)

Almost adverb Very nearly	Nearly adverb 1. In a near manner; closely; intimately 2. Particularly 3. Almost, all but

Figure 4. Two (native-speaker) monolingual dictionary entries for *almost* and *nearly*

It is also the case with two well-known dictionaries for learners of English:
The examples they give can be quite meaningful to those who have become experts in the difference between the two adverbs, but anyone else will have to be satisfied with circular definitions which will not help them find underlying rules of use.

Cambridge Advanced Learner's Dictionary (2003)

Almost	Nearly
adverb nearly: She's almost thirty. It was almost six o'clock when he left. I almost wish I hadn't invited him. It'll cost almost as much to repair it as it would to buy a new one. Almost **all** the passengers on the ferry were French. They'll almost **certainly** forget to do it. The town was almost entirely destroyed during the war. We were bitten by mosquitoes almost **every** night. The boat sank almost **immediately** after it had struck the rock. Most artists find it almost **impossible** to make a living from art alone.	adverb almost, or not completely: It's been nearly three months since my last haircut. I've nearly finished that book you lent me. She's nearly as tall as her father now. They'd eaten nearly everything before we arrived. FIGURATIVE It was so funny - we nearly died laughing.

Longman Dictionary of Contemporary English (2003)

Almost	Nearly
adverb Nearly, but not completely or not quite Have you almost finished? Supper's almost ready. It was almost midnight Almost nothing was done to improve the situation The story is **almost certainly** true He's **almost as** old **as** I am **Almost all/every/everything** Marsha visits her son almost every day.	adverb 1. especially British English almost, but not quite or not completely; **almost**: it took nearly two hours to get here. Michelle's nearly twenty. Is the job nearly finished? Louise is nearly as tall as her mother. I **nearly always** go home for lunch. He **very nearly** died. 2. **not nearly** not at all: He's not nearly as good-looking as his brother. We've saved some money, but it's **not nearly enough**.

Figure 5. Two learner dictionary entries for *almost* and *nearly*

It seems that dictionaries, even learners' ones, tend to define and show usage, but they are not very good at distinguishing so-called synonyms, unless there is a separate usage box at the end or on the CD that comes with them. For example, the CD accompanying the Longman dictionary of contemporary English had this extra information, which is an improvement on mere synonymy but is still not enough:

> almost/nearly:
> [adverb] use this to say that something is a little less than a number or amount, or a little before a particular time. **Almost** and **nearly** have the same meaning, but **almost** is much more common than **nearly** in American English. In British English both words are common.

Even bilingual French-English dictionaries which are often used by French students and teachers see *almost* and *nearly* as equal synonyms of *presque*.
Le petit Robert bilingue (1987)

Presque	Almost	Nearly
Adv. a) Almost, nearly, virtually b) Contexte négatif: hardly, scarcely, almost, virtually	Adv. Presque	Adv. a) (almost) Presque, à peu près, près de b) Not nearly: loin de c) (closely) près, de près

Larousse (1995)

Presque	Almost	Nearly
Adv. 1. Dans phrases aff. Almost, nearly 2. Dans phrases neg. Hardly, almost 3. Quasi = almost, nearly, practically	Adv. Presque	Adv. [almost] Presque, à peu près

Figure 6. Two bilingual dictionary entries for *almost* and *nearly*

The recurrent use of circularity (*almost* = *nearly* = *almost*) hides differences between *almost* and *nearly* which are not covered by a simple explanation. It is more a matter of nuances, some of which do appear in dictionaries, but only in the form of examples and never explicitly enough for teachers or students to teach or learn a definite rule. If the dictionaries quoted here are the main references, for years to come, *almost* and *nearly* will still be presented as synonyms.

2 A corpus approach

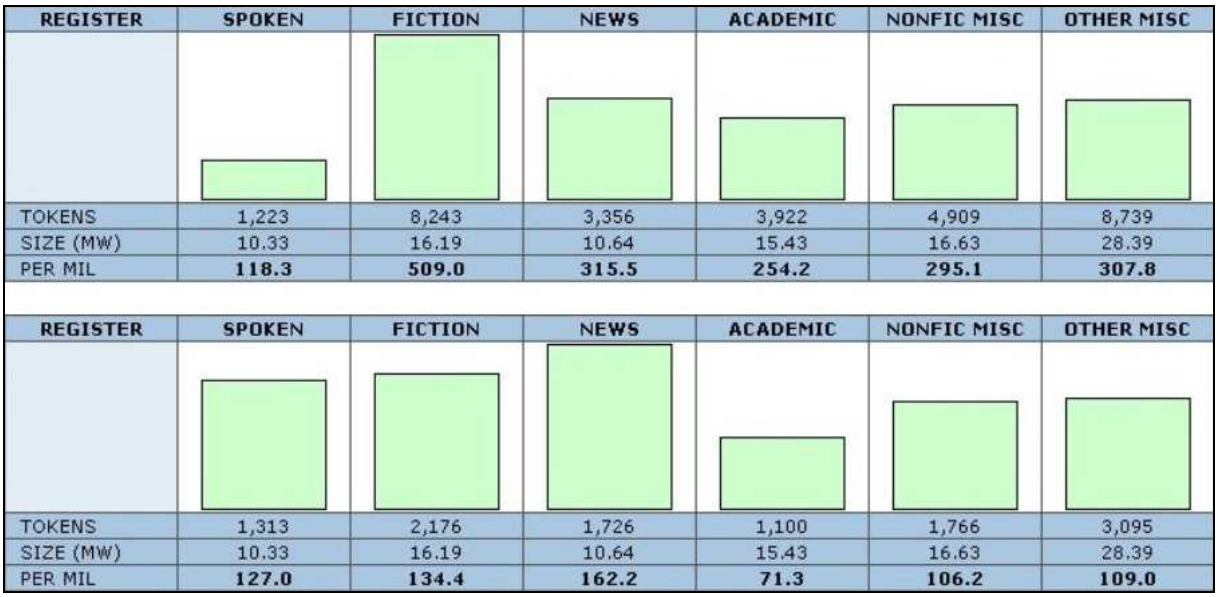
It is a fairly basic precept of linguistics that total synonymy is rare, if it exists at all, yet we have seen that intuitions and traditional references sources are of little help in establishing useful differences between apparently synonymous items such as *nearly* and *almost*. However, Kjellmer (2003b) does manage to do just that, finding differences in terms of frequency, distribution in different registers, coverage, collocation and colligation. He does this by using a large corpus of contemporary English, in this case part of the COBUILD Bank of English (henceforth BoE). There are a number of advantages to a corpus-based approach to language study, not the least of which is that it allows us to refer to multiple occurrences of language in use and hence "to access the combined intuitions of literally thousands of native speakers together" (Frankenberg-Garcia, 2005b: 192). In other words, corpora offer a way to combine reference and intuition in a single large resource.

The rest of this paper attempts to show how such a corpus approach can be applied to *almost* and *nearly*, as it represents a typical example of "one of the commonest types of question asked by the enquiring learner", namely 'What is the difference between…?' (Johns, 1991a: 4). This is the type of question where corpora can be of considerable help to specialist and non-specialist alike, and provide a valuable complement to traditional resources. In an attempt to show how teachers and learners might benefit from such tools, we shall be using only simple techniques based on highly user-friendly software found free on the internet.

The BoE mentioned above is an expensive corpus, and only the simplest searches of a small subcorpus can be conducted free[1]. The British National Corpus (henceforth BNC) is much cheaper, and although the official demonstration interface[2] is limited in ways similar to that of the BoE, relatively sophisticated research of the entire corpus can be conducted using an interface designed by Mark Davies of Brigham Young University, called Variation in English Words and Phrases (VIEW[3]). Briefly, the BNC contains 100 million words of British English gathered in the early 1990s, 10% of which is transcribed speech. While much research in corpora and language teaching/learning emphasise the use of small, tailor-made corpora, a large, principled reference corpus such as the BNC offers a wider view of the language as a whole, especially useful when different registers can be searched individually (see Aston, 1997, for a summary of the arguments).

2.1. Frequency and register

A first stage is simply to input *almost*/*nearly* into the "search string" box at the top left of the screen to obtain the overall frequency for each item. This shows that *almost* occurs 30,392 times in the entire 100 million word corpus, i.e. 303.92 times per million words (pmw), while *nearly* occurs 11,176 times (111.76 pmw). Overall then, *almost* is 2.72 times as frequent as *almost*. By selecting the "chart" option in the display section we can compare the distribution over six "macro-registers" (figure 7).



REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
TOKENS	1,223	8,243	3,356	3,922	4,909	8,739
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**118.3**	**509.0**	**315.5**	**254.2**	**295.1**	**307.8**

REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
TOKENS	1,313	2,176	1,726	1,100	1,766	3,095
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**127.0**	**134.4**	**162.2**	**71.3**	**106.2**	**109.0**

Figure 7. Distribution in macro-registers for *almost* (top) and *nearly* (bottom).

The display is revealing, suggesting for example that *almost* is more typical of literary registers. Unfortunately, it does not take into account the difference in overall frequency of each item, but this can be offset by copying the figures into a spreadsheet (figure 8). The two miscellaneous registers are omitted as they are close to what we would expect, but the other

[1] http://www.collins.co.uk/Corpus/CorpusSearch.aspx, last consulted August 2007.
[2] http://www.natcorp.ox.ac.uk, last consulted August 2007.
[3] http://view.byu.edu, last consulted August 2007. The site is accompanied by a "three-minute tour" and many help pages.

four main registers show significant differences. Both *almost* and *nearly* occur less frequently in the Academic register than in the corpus as a whole, and both are overrepresented to different degrees in the News and Fiction registers; this presumably reflects differing degrees of emphasis on precision or approximation. More specifically, *almost* is comparatively more likely to be found in the News and Academic registers, while *nearly* is comparatively more frequent in Fiction. Only in the Spoken register do the norms differ in opposite directions: *nearly* is thus more a spoken item than *almost*, a finding which conflicts with Kjellmer (who found that "neither of them is used much in the spoken language;" 2003: 26), and is thus worth pursuing.

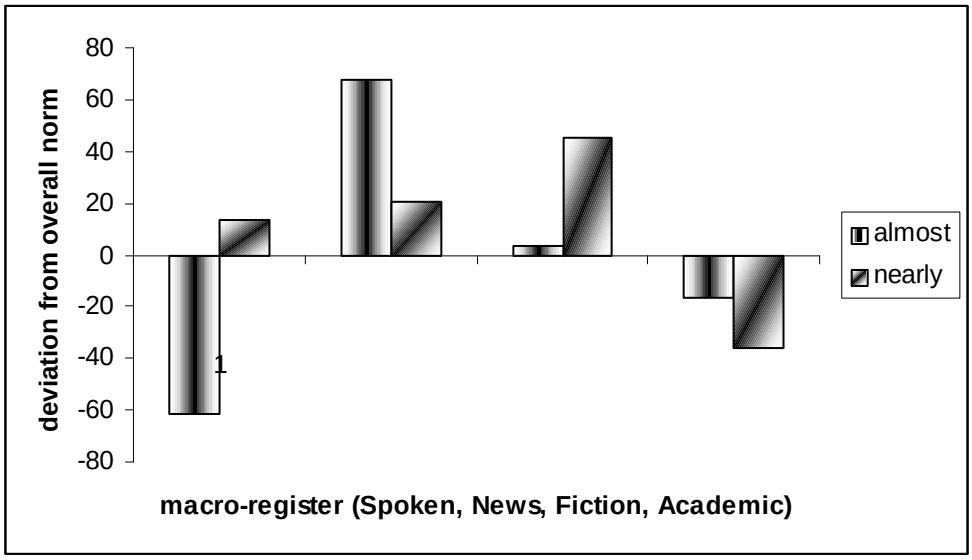
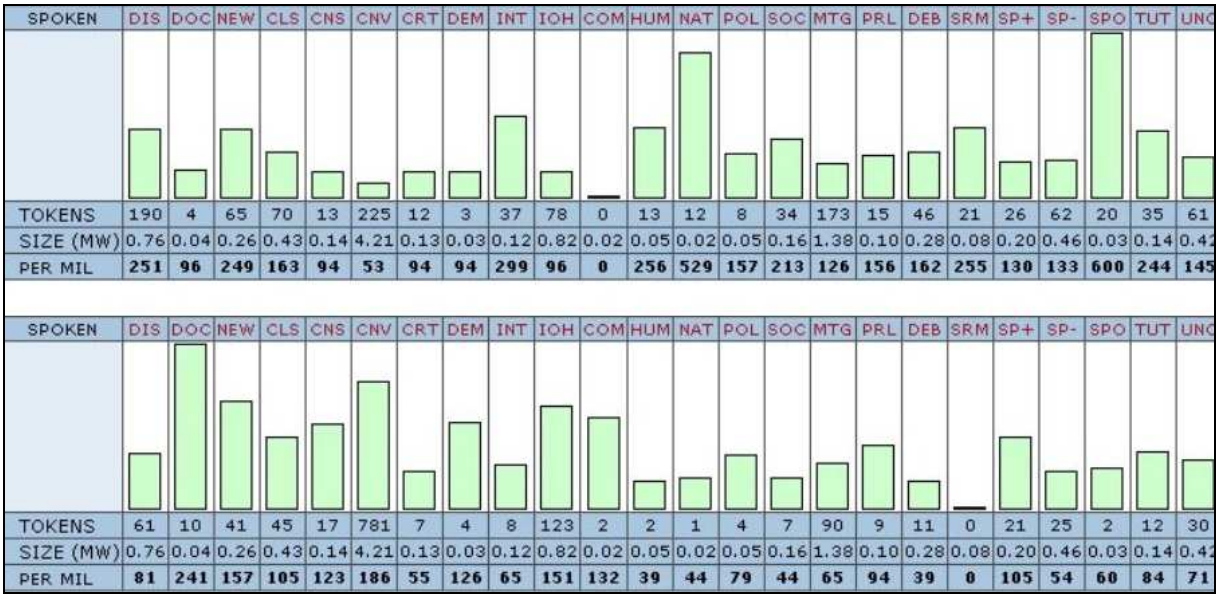


Figure 8. Distribution by macro-register.

By clicking on the column for the Spoken register, VIEW automatically subdivides it into 24 micro-registers, from Sermons to Interviews (figure 9). The small and variable size of each micro-register makes serious comparisons difficult: the largest at over four million words is CNV Conversation, where the difference is quite striking. Of the 24 spoken registers, *almost* is extremely underrepresented in Conversation (only COM Lectures Commerce uses it less frequently), while *nearly* is extremely overrepresented (only DOC Broadcast Documentaries use it more frequently). It is possible that Kjellmer (2003) did not find any correlation with spoken language simply because his software did not allow such register-sensitive queries.



SPOKEN	DIS	DOC	NEW	CLS	CNS	CNV	CRT	DEM	INT	IOH	COM	HUM	NAT	POL	SOC	MTG	PRL	DEB	SRM	SP+	SP-	SPO	TUT	UNC
TOKENS	190	4	65	70	13	225	12	3	37	78	0	13	12	8	34	173	15	46	21	26	62	20	35	61
SIZE (MW)	0.76	0.04	0.26	0.43	0.14	4.21	0.13	0.03	0.12	0.82	0.02	0.05	0.02	0.05	0.16	1.38	0.10	0.28	0.08	0.20	0.46	0.03	0.14	0.42
PER MIL	**251**	**96**	**249**	**163**	**94**	**53**	**94**	**94**	**299**	**96**	**0**	**256**	**529**	**157**	**213**	**126**	**156**	**162**	**255**	**130**	**133**	**600**	**244**	**145**

SPOKEN	DIS	DOC	NEW	CLS	CNS	CNV	CRT	DEM	INT	IOH	COM	HUM	NAT	POL	SOC	MTG	PRL	DEB	SRM	SP+	SP-	SPO	TUT	UNC
TOKENS	61	10	41	45	17	781	7	4	8	123	2	2	1	4	7	90	9	11	0	21	25	2	12	30
SIZE (MW)	0.76	0.04	0.26	0.43	0.14	4.21	0.13	0.03	0.12	0.82	0.02	0.05	0.02	0.05	0.16	1.38	0.10	0.28	0.08	0.20	0.46	0.03	0.14	0.42
PER MIL	**81**	**241**	**157**	**105**	**123**	**186**	**55**	**126**	**65**	**151**	**132**	**39**	**44**	**79**	**44**	**65**	**94**	**39**	**0**	**105**	**54**	**60**	**84**	**71**

Figure 9. Distribution in micro-registers for *almost* (top) and *nearly* (bottom)

Key: 1) DIS = broadcast discussion; 2) DOC = broadcast documentary; 3) NEW = broadcast news; 4) CLS = classroom; 5) CNS = consultation; 6) CNV = conversation; 7) CRT = courtroom; 8) DEM = demonstration; 9) INT = interview; 10) IOH = interview oral history; 11) COM = lecture commerce; 12) HUM = lecture humanities arts; 13) NAT = lecture natural science; 14) POL = lecture politics law education; 15) SOC = lecture social science; 16) MTG = meeting; 17) PRL = parliament; 18) DEB = public debate; 19) SRM = sermon; 20) SP+ = speech scripted; 21) SP- = speech unscripted; 22) SPO = sports live; 23) TUT = tutorial; 24) UNC = unclassified.

Such statistics are virtually instantaneous with most corpus software, and allow us to see at a glance that differences in frequency and distribution are common between items such as these.

2.2. Concordance data

One of the most common classroom techniques is to provide a concordance for learners to examine. This can be done simply with VIEW by returning to the “table” option in the display section, entering the search string (*almost/nearly*) and clicking on one or other word (an example is given in the appendix). Although only 100 lines are presented on a single page, the concordance thus generated can often be quite revealing and allow learners to uncover specific patterns of use. In the present case, it is immediately obvious that *nearly* occurs far more frequently with numerals and other quantitative expressions than *almost* – 28 compared to only 10 (figure 10).

almost 1,000 death sentences, of which at least 750 resulted in executions.
almost 100 drawn from different classes, sizes and styles of hotels geographically
almost 20 years; Ahmad "Abd al-Rau'uf Roummou, a 55-year-old teacher
almost 80 other prisoners freed at the same time were held for political
almost 80% of total turnover in the fast food market. In the
almost a quarter of a century. "Which," continued his
almost half (46%) of AIDS related deaths. While pneumonia continues
almost half (46%) of those with AIDS in 1986. While
almost one meal in every 10 served in the UK. The boom
almost two-thirds of UK fast food units were franchised. The report forecasts

nearly £½million, Casey Jones was obviously going to pay for himself only
nearly 250 years later would have never taken place --; and my humble

nearly 3,000 have already died. 1 in 500 Londoners are believed to
nearly 30 capsicum varieties are listed in seed catalogues. Colours range from
Nearly 30 years after writing it Leonard was enthusiastically recommending our attendance at
nearly 30 years to compile the information on these machines which began in
nearly 300 Ugandan nationals currently detained for political reasons in Rwanda. They
nearly 4 years like Steffi? Well I rest my case on that
nearly 5,000 pubs that serve top-notch cask beer --; there are more than 1,500
nearly 5,000 reported cases of nearly 3,000 have already
nearly 60% of pensioners receive at least 75% of their income from state benefits
nearly 7 million sold on stations in 1985 --; lost their "curly
nearly 80 special screening centres throughout England --; at hospitals, clinics or
nearly a third of tourism expenditure by Britons in this country. During
nearly a thousand feet up, and I could see that the gladioli
nearly fifty-four, aren't you? I don't mean to be
nearly four million square miles, almost half of which constitutes the backbone
nearly half a year between my interview and my starting there passed without
nearly half of all the pasta eaten in Britain is the canned variety
nearly half the Class 37 fleet, with a number of variants designed
nearly half the pubs which were standing in the 1950s, including some
nearly nine years ago now …" The conversation continued for a couple
nearly seven o'clock. The news would be starting on Channel Four.
nearly ten years ago. I've done more normal jobs since then
nearly thirty years before then it had displayed the result of my own
nearly thirty years on, lighting a cigarette, tossing away a cloud
nearly thirty years, incurred a sequel, in which the discussion,
nearly thirty years, the first AC electric Co-Co locomotive arrived in 1986

Figure 10. Immediate right collocates: numerals in the first 100 lines of VIEW

2.3. Collocates

Such an approach can lead to useful insights, but they do need checking. This can be done quite simply by entering *almost*/*nearly* in the search window followed by the tag for cardinal numbers [CRD] from the dropdown menu just beneath. The top 10 hits with *almost* occur 8.66 pmw, while the top 10 with *nearly* occur 14.39 pmw. In other words, even without allowing for the fact that *nearly* is about three times less frequent in the BNC than *almost*, it does indeed seem to occur considerably more frequently before a cardinal, thus confirming our initial concordance finding.

A more rigorous approach is to get the software to look for collocates. This can be done for each word individually, or by entering *almost/nearly*, then checking “surrounding” in the display menu and, for the immediate right collocate (R1), with 0 words to the left and 1 word to the right (figure 11). This shows that, for example, *certainly* occurs 1,494 times in the corpus after *almost*, but never after *nearly*. The strongest right collocate for *nearly* is *all*, which is also a strong right collocate for *almost*.

	ALMOST +	# ALMOST	# NEARLY	% ALMOST	MI SCORE		NEARLY +	# NEARLY	# ALMOST	% NEARLY	MI SCORE
1	CERTAINLY	1494	0	100.0%	5.60	1	ALL	1816	1320	40.8%	3.46
2	ALL	1320	908	59.2%	2.84	2	A	1100	1029	34.8%	0.83
3	A	1029	550	65.2%	0.46	3	TWO	788	215	64.7%	3.14
4	AS	819	336	70.9%	1.64	4	ALWAYS	764	480	44.3%	4.34
5	EVERY	788	246	76.2%	4.20	5	HALF	684	347	49.6%	4.66
6	ENTIRELY	595	1	99.8%	5.67	6	AS	672	819	29.1%	1.75
7	THE	552	105	84.0%	-1.20	7	EVERY	492	788	23.8%	4.03
8	IMPOSSIBLE	490	27	94.8%	5.46	8	THREE	474	142	62.5%	3.30
9	ALWAYS	480	382	55.7%	3.57	9	FOUR	310	67	69.8%	3.42
10	ANY	420	5	98.8%	2.43	10	FIVE	246	75	62.1%	3.31

Figure 11. Top 10 exclusive R1 collocates for *almost* and *nearly* (by frequency)

The same search, but this time sorting by "percent", lists the right collocates which occur most frequently for each item and least frequently for the other (figure 12). *Certainly* is still of course highest on the list for *almost*, which has 97 separate words which occur 12 times or more in the BNC in the R1 position but never with *nearly*. On the other hand, there are only seven words which occur 10 times or more in this position with *nearly* but never with *almost*. Top of the list is *man* with a frequency of 20, mostly in the expression *the nearly man* in political contexts, as can be seen by clicking on *man* (figure 13). All of this suggests that *nearly* is more limited in its unique collocates, *almost* has wider coverage.

	ALMOST +	# ALMOST	# NEARLY	% ALMOST	MI SCORE		NEARLY +	# NEARLY	# ALMOST	% NEARLY	MI SCORE
1	CERTAINLY	1494	0	100.0%	5.60	1	MAN	20	0	100.0%	0.42
2	IMMEDIATELY	395	0	100.0%	4.85	2	CHRISTMAS	16	0	100.0%	2.12
3	AS IF	314	0	100.0%	4.28	3	120	12	0	100.0%	0.00
4	EXCLUSIVELY	304	0	100.0%	6.39	4	2.5	12	0	100.0%	0.00
5	INVARIABLY	207	0	100.0%	6.10	5	CENTRAL	12	0	100.0%	1.03
6	ANYTHING	193	0	100.0%	3.15	6	DINNER	12	0	100.0%	2.21
7	NO	189	0	100.0%	1.12	7	1.5	10	0	100.0%	0.00
8	INEVITABLY	98	0	100.0%	4.66	8	&POUND;7	8	0	100.0%	0.00
9	AS THOUGH	93	0	100.0%	4.12	9	37	8	0	100.0%	0.00
10	INEVITABLE	90	0	100.0%	4.69	10	BLOWN	8	0	100.0%	3.39

Figure 12. Top 10 exclusive R1 collocates for *almost* and *nearly* (by per cent)

it would have been this one." The Cabinet Reshuffle	**Nearly Man** is there at last THERE was a fear that John Patten might never
time coming, prompting some commentators to dub him "The	**Nearly Man**". When Lord Waddington moved to the Home Office, Mr
Telegraph, reports … ELLCOCK'S SAD SURRENDER. "	**NEARLY man**" is an overused term in sport, a label that is
." Soon afterwards, however, Ellcock became the "	**nearly man**" again, entering hospital for further surgery to remove the
to the Melbourne Sunday Herald-Sun. Neil Mallender, the "	**nearly man**" Somerset seamer Neil Mallender was called up as a standby for
has so far eluded him. Derick Allsop profiles Britain's	**nearly man** and assesses the others in the running The face is florid and
is the year Carlos Cardus must prove he is not the	**nearly man** of 250 GP racing. A late developer on the world championship
Championship in 1983 and 1984, Alain Prost lose the "	**nearly man** " tag in 1985 as he became the first Frenchman to win
beginning of his recovery to the kindly determined man, or	**nearly man**, he had been before Sobibor. These things he barely understood
really got past the stage of being regarded as a "	**nearly man**" a jockey who never quite clicked with the racing public.

Figure 13. First 10 concordance lines for *nearly man*

Looking at the lists in more detail brings to light a number of other tendencies in the types of words which occur as the immediate right collocates. In particular, many of the most frequent with *almost* are adverbs. To take this further, we can resubmit the same query but this time selecting adverb [adv.ALL] in the drop-down menu in the display section. Figure 14 is immediately revealing on a number of counts. Firstly, the collocates are far stronger for *almost*, with the top 10 occurring a total of 2,730 times, while those for *nearly* occur only 21

times in the entire corpus. Secondly, the adverbs which occur exclusively with *almost* overwhelmingly end in -LY, whereas those with *nearly* constitute a far more heterogeneous group. This is not to say that the string *nearly *ly* is impossible[4] (there are 30 occurrences in the entire BNC), but it is unusual, perhaps for purely phonetic reasons.

	ALMOST +	# ALMOST	# NEARLY	% ALMOST	MI SCORE		NEARLY +	# NEARLY	# ALMOST	% NEARLY	MI SCORE
1	CERTAINLY	1494	0	100.0%	5.60	1	FAR	4	0	100.0%	-0.03
2	IMMEDIATELY	395	0	100.0%	4.85	2	NOW	3	0	100.0%	-1.61
3	EXCLUSIVELY	304	0	100.0%	6.39	3	THREEFOLD	2	0	100.0%	4.78
4	INVARIABLY	207	0	100.0%	6.10	4	THREE-FOLD	2	0	100.0%	5.86
5	INEVITABLY	98	0	100.0%	4.66	5	PROPORTIONALLY	2	0	100.0%	5.31
6	UNIVERSALLY	80	0	100.0%	6.12	6	LONG	2	0	100.0%	-1.05
7	IMPERCEPTIBLY	43	0	100.0%	7.11	7	FUCKING	2	0	100.0%	1.76
8	OVERNIGHT	40	0	100.0%	4.23	8	EVER	2	0	100.0%	-0.32
9	INSTANTLY	38	0	100.0%	4.41	9	FOR EVER	1	0	100.0%	1.97
10	DAILY	31	0	100.0%	2.61	10	HALF-WAY	1	0	100.0%	2.99

Figure 14. Top 10 exclusive R1 adverb collocates for *almost* and *nearly*

Similar queries can be conducted for the immediate left collocate (figure 15). Unsurprisingly, the most frequent exclusive L1 collocate of *almost* is *an*. It seems however that *nearly* is more easily modified, with *most* and *pretty* top of the list; whereas ?*most almost* might not occur in this corpus of British English for phonetic reasons, it is more difficult to find a convincing explanation for the non-occurrence of ?*pretty almost*.

	ALMOST +	# ALMOST	# NEARLY	% ALMOST	MI SCORE		NEARLY +	# NEARLY	# ALMOST	% NEARLY	MI SCORE
1	AN	956	0	100.0%	-0.07	1	MOST	60	0	100.0%	1.02
2	SEEMED	124	0	100.0%	0.61	2	PRETTY	18	0	100.0%	2.37
3	SOUNDED	31	0	100.0%	1.33	3	YEAH	16	0	100.0%	-0.15
4	FOLLOWED	18	0	100.0%	-0.89	4	TILL	10	0	100.0%	2.13
5	'LL	15	0	100.0%	0.00	5	ALL	8	0	100.0%	-1.96
6	CONCERNED	14	0	100.0%	-1.25	6	CLAIMED	8	0	100.0%	0.78
7	CONSISTED	13	0	100.0%	1.18	7	AFFECTING	6	0	100.0%	2.74
8	LOOKING	13	0	100.0%	-1.78	8	AWAKE	6	0	100.0%	3.04
9	SEEMED	12	0	100.0%	-1.72	9	COLLECT	6	0	100.0%	2.27
10	WAY	12	0	100.0%	-3.16	10	EQUALLED	6	0	100.0%	5.00

Figure 15. Top 10 exclusive L1 adverb collocates for *almost* and *nearly*

More striking still is the prevalence of verbs of appearance left of *almost*: *seemed*, *sounded*, *looking*, and so on. Clicking on the first of these produces a concordance which shows that the string *seemed almost* is typically followed by an adjective or a verb (figure 16). As with adverbs, these parts of speech are also more frequent R1 collocates of *almost*, which thus seems to serve more readily as a hedging device.

avid for every shilling he could earn. And he had	**seemed almost**	to be currying favour when he was tumbling out the story of
. Soft-centred milk chocolates that came as one bit them	**seemed almost**	as exciting as dancing with someone you knew you were
mountains was still taking place now that the man who had	**seemed almost**	to stand for its yearly renewal had gone. I'll tell you
a decade later, when the term "has-been"	**seemed almost**	an understatement, she not only gratefully accepted but
of the press. Journalists themselves as well as the party	**seemed almost**	oblivious of the harm being wrought. The third congress of
Until a sudden decline from September 1 978, he	**seemed almost**	unchallengeable. He was fundamentally assisted by political
and a considerable swathe of the people of Scotland,	**seemed almost**	to contract out from traditional forms of participation and civic

[4] It is general practice in many corpus programs to use the asterisk (*) as a wild card representing any word or any group of letters.

such rigour, both in economic and political terms,	**seemed almost**	unacceptable. In fact, the government's policy from 1982
suffered accordingly. By the end of 1981, she	**seemed almost**	a Prime Minister at bay. Throughout 1980 she had battled
. "You must be insane." The doctor	**seemed almost**	embarrassed by her own outburst, but Phoebe was relieved. It

Figure 16. Sample 10 concordances for *seemed almost*

2.4. Data and interpretation

Conclusions such as these take us from the mere facts of corpus data and into the realms of interpretation. It is frequently remarked that corpus linguistics is a "methodology" (e.g. McEnery *et al*., 2006: 8), and thus just another tool in the linguist's armoury, albeit an extremely powerful one. Indeed, facts alone are of little use unless they can be interpreted, and corpus linguistics is at its strongest when allied with other methods of analysis. An example of this can be seen in a recent series of postings on Language Log on one aspect of *almost* and *nearly*. The first, by Mark Liberman[5], uses corpus data to test the intuition that "*nearly* tends to be uneasy when asked to modify overtly negative words like *no*, *never* and *none*." A number of alternative hypotheses are ruled out, e.g. that *nearly* is more "positive", as *almost* is more frequently found with *everyone* as well as with *no one*. The follow up[6] also rules out differences of "concrete" vs. "abstract" usage, as well as avoidance of alliteration, since *nearly nine* is found more frequently than *almost nine* (especially remembering that, overall, *almost* is considerably more frequent than *nearly*). One proposal is that *nearly* means "a number slightly less than", which is refined in a guest posting by Jerry Sadock[7] in terms of conversational implicature, which suggests that "*nearly n* exceeds (hence is better than) what was expected or hoped for, while *almost n* does not conventionally connote any particular desire, hope or expectation." This would make *nearly* unusual when followed by concepts such as *nothing*, *none* or *nobody* which rarely "exceed expectations" – although contexts can be found which can explain the exceptions. A final guest posting by Lucia Pozzan and Susan Schweitzer[8] reverses the situation and looks at negatives before *nearly* or *almost*; *not nearly* is far more common than *not almost* (which is likely to be confined to echo contexts) and of course means not just *not* + *nearly* but something like "a long way from", and is highly loaded affectively. These discussions show that such questions do preoccupy linguists today, and most arguments are supported by corpus data of some kind.

2.5. Summary

As Kjellmer (2003b: 26) concludes in his article on *almost* and *nearly*, the power of corpus linguistics allows us to show that "far from being the next-to-interchangeable synonyms that dictionaries could lead us to imagine, *almost* and *nearly* turn out, on closer inspection, to be partly overlapping but in important respects clearly contrasting words." Our own brief analysis of these items in the BNC has highlighted the following aspects:

a) Both are highly frequent items, but *almost* is just under three times more frequent than *nearly*.
b) Distribution varies according to register, with *almost* being more typical of literary registers, and *nearly* occurring more frequently in speech.

---

[5] 14/06/07. http://itre.cis.upenn.edu/%7Emyl/languagelog/archives/004604.html, last consulted August 2007.
[6] 16/06/07. http://itre.cis.upenn.edu/~myl/languagelog/archives/004613.html#more, last consulted August 2007.
[7] 24/06/07. http://itre.cis.upenn.edu/~myl/languagelog/archives/004640.html#more, last consulted August 2007.
[8] 25/07/07. http://itre.cis.upenn.edu/~myl/languagelog/archives/004705.html#more, last consulted August 2007.

c) *Almost* has wider coverage than *nearly*, which tends to have a more restricted range of collocates.
d) L1 collocates for *almost* include many verbs of appearance, suggesting its use as a hedging device, while *nearly* can be modified more easily with items such as *most* or *pretty*.
e) R1 collocates for *almost* are typically adverbs (especially in –LY) and adjectives, whereas *nearly* is more likely to modify a noun, most strikingly numbers.

Although these may not, with hindsight, seem particularly striking, Louw (1997: 250) is impatient with the "I knew it all along" phenomenon in corpus linguistics, and suggests that "instances of the phenomenon of '20:20 hindsight'… ought to be dealt with very firmly indeed."

There is not the space here to follow up all aspects of use of even these two items, but the availability of VIEW means that these findings can be repeated, the results checked with modified queries, and further searches conducted – all using the same corpus. It should be stressed that the findings can only be as good as the corpus they come from; the non-occurrence of a particular string is not evidence that it is impossible, merely that it is not attested in that corpus. Specifically, our results only apply to British English from the late 20th century; they tell us nothing about the evolution of the language, or about other varieties such as American English[9].

While our study of *almost* and *nearly* in the BNC has allowed us to uncover some useful information about these items, it is clear that the information itself is qualitatively different from that typically sought in most dictionaries and traditional reference sources. In general, most users (whether natives or non-natives) tend to look for hard-and-fast rules, while corpus data generally delivers tendencies and patterns of usage.

2.6. Corpora in dictionaries

There seems little reason why dictionaries should not incorporate corpus findings, and even corpus techniques in electronic dictionaries (Cobb, 2003a). Indeed, publishers have been quick to see the potential of corpora in lexicography, to the extent that today it would be virtually unthinkable to embark upon a major dictionary of any kind without recourse to corpus linguistics. All of the recent dictionaries presented in the introduction claim to be corpus-based, although this is, as we have seen, no guarantee that important differences will be highlighted. One popular dictionary which has worked more thoroughly on *almost* and *nearly*, however, is the Oxford Advanced Learner's Dictionary (Hornby & Wehmeier, 2007[10]). In an on-line article[11] the current editor explicitly notes that "dictionary writers today… have huge electronic databases with many millions of words in them, that they can consult to see how many different speakers have used a word in a whole range of situations,"

---

[9] The American National Corpus, a partial parallel of the BNC, is currently under construction; see http://www.americannationalcorpus.org/ (last consulted August 2007). Existing corpora of American English tend to be rather small or outdated, or less carefully balanced than the BNC. Partly for such reasons, many linguists use the internet as an immediate reference source, and many tools to extract linguistic data from the web currently exist or are being developed.
[10] The OALD can be consulted free on line at http://www.oup.com/elt/catalogue/teachersites/oald7/lookup?cc=global, last consulted August 2007.
[11] http://www.oup.com/elt/catalogue/teachersites/oald7/about_OALD/word_from_editor?cc=global, last accessed August 2007.

and refers the reader to an OALD article on the BNC which briefly describes the use of concordances[12]. The use of corpora is no guarantee that definitions will not be circular, however:

- *almost* = not quite; SYN. nearly
- *nearly* = almost; not quite; not completely

Nevertheless, the influence of corpora is evident in the accompanying usage box (figure 17), which contains information on frequent collocates and distribution in different registers.

SYNONYMS

almost · nearly · practically

These three words have similar meanings and are used frequently with the following words:

almost ~	nearly ~	practically ~
certainly	(numbers)	all
all	all	every
every	always	no
entirely	every	nothing
impossible	finished	impossible
empty	died	anything

- They are used in positive sentences: *She almost/nearly/practically missed her train.* They can be used before words like *all*, *every* and *everybody*: *Nearly all the students have bikes.* ◇ *I've got practically every CD they've made.* **Practically** is used more in spoken than in written English. **Nearly** is the most common with numbers: *There were nearly 200 people at the meeting.* They can also be used in negative sentences but it is more common to make a positive sentence with **only just**: *We only just got there in time.* (or: *We almost/nearly didn't get there in time.*)
- **Almost** and **practically** can be used before words like *any, anybody, anything*, etc.: *I'll eat almost anything.* You can also use them before *no, nobody, never*, etc. but it is much more common to use **hardly** or **scarcely** with *any, anybody, ever*, etc.: *She's hardly ever in* (or: *She's almost never in*).
- **Almost** can be used when you are saying that one thing is similar to another: *The boat looked almost like a toy.*
- In *BrE* you can use *very* and *so* before **nearly**: *He was very nearly caught.*

Figure 17. OALD usage box for *almost, nearly* and *practically*

The presence of numerous sentence-level examples is a valuable aid, although the choice of examples in the main entries is not perhaps uncontroversial when compared against the BNC. Amongst other things, similar examples elicit no comment; the usage box gives *almost all* as well as *nearly all*, with examples "I like *almost all* of them" and "The audience was *nearly all* men" (emphasis added). The relatively minor differences in frequency between *almost all (of)* and *nearly all (of)* might explain this, but the same cannot be said of "It's almost time to go" and "It's nearly time to leave": *nearly time* and *nearly time to* are three and six times more frequent respectively — particularly remarkable remembering that *nearly* is three times less frequent in the corpus as a whole. Other examples include "It's a mistake they *almost always* make" and "They're *nearly always* late" (emphasis added); while *almost always* is only slightly more frequent than *nearly always*, it might be useful to learn that *nearly always* is three times more common in speech. Finally, the examples given are of common usage, but there is little indication of highly unusual collocates; the dictionary user is left to infer that *nearly* –LY is rare, for example.

Conclusion

[12] http://www.oup.com/elt/catalogue/teachersites/oald7/more_on_dicts/bnc?cc=global, last consulted August 2007.

In this paper we have attempted to show that intuitions are often insufficient to answer typical questions learners may have about language, and even traditional references sources do not always provide satisfying answers. Using the example of a pair of near-synonyms, *almost* and *nearly*, and a single large corpus which can be accessed free on the internet, we have attempted to demonstrate how corpus linguistics can provide one possible solution. The kinds of queries shown here require no particular expertise, and should be well within the grasp of most teachers. In addition to checking language points as outlined here, corpora also provide a ready source of authentic co-texts for language teaching and testing materials.

Owen (1996: 219) has pointed out that the role of teachers in many cultures is to be an absolute expert, "so that doubt or hesitation in delivering judgement is normally taken for ignorance rather than wisdom." Corpus studies have highlighted the intrinsically "fuzzy" nature of language, which may appear as chaos or disorder to many learners (and teachers) even at very advanced levels. A corpus approach here can help by showing that patterns and tendencies contribute to creating order in this chaos, while rigid adherence to rules is likely to lead to oversimplification or unnecessary levels of abstraction. Regular reference to corpora to answer students' questions, either in the classroom or between sessions, would seem preferable to the classic response of "we just don't say that" — surely a worse admission of ignorance. For non-native teachers in particular, corpora can be immensely empowering for just this reason.

Johns (e.g. 1991b) has frequently made the case that teachers can use corpora directly with students, culminating in his "kibbitzing" approach (Johns, 1997b)[13]. While even lower level learners may derive some benefits (Boulton, 2009, in press), they are perhaps most suitable for advanced or specialist language students, who can be encouraged to use them on their own for many questions where they traditionally open a dictionary or grammar book. Training will certainly increase the benefits (just as dictionary training can bring rewards), but "the difficulties [of using corpora] should not be overestimated; learners should quickly acquire the skills needed" (Bernardini, 2001: 243). Even without training, "both teacher and student can make use of a corpus right away, with only a modest few hours of orientation" (Sinclair 2004b: 288). In other words, it is no more necessary to be a corpus linguist to derive benefit from this approach than it is to be a lexicographer to use a dictionary.

The potential uses of corpora in language learning are, as Breyer (2006: 162) puts it, "limited only by the imagination of the user." They can be an aid to understanding of other-produced texts, and more obviously as a reference aid in many production tasks. In particular, their use is ubiquitous on most contemporary translation courses, and greatly contributes to learner autonomy. We have seen how they can help in understanding nuances not clearly explained in traditional resources, and they can certainly be used for learning the language. Clearly learners can scarcely be expected to go to the lengths outlined here for every item they want to learn, but the very fact of using corpora in this way is likely to lead to wider benefits. Allan (2006), for example, found that students who used corpora improved their learning of the target vocabulary and other items as well. It seems likely then that the process of discovery, of thinking about the language, of formulating queries, scanning concordances and interpreting results can all help to improve noticing and other language skills essential for long-term learning. And, of course, corpora can be used for the "big themes" of language learning such as grammar, not just usage details of specific lexical items (Boulton, 2007).

While it is not possible to cover all aspects here, many excellent sources of ideas already exist, of which the following are just a few. Biber *et al*. (1999) and McEnery *et al*. (2006) provide clear insights into the methodology of corpus linguistics, Hunston (2002) with

[13] Dozens of examples are illustrated at http://www.eisu2.bham.ac.uk/johnstf/timeap3.htm#revision, last consulted August 2007.

a focus more explicitly on applied linguistics. Burnard and McEnery (2000), Aston (2001) and Sinclair (2004a) are inspiring collections of papers on the uses of corpora in language teaching and learning. More concrete applications can be found in Thurston and Candlin (1997), Tribble and Jones (1997) and O'Keeffe *et al*. (2007). Finally, an awareness of corpus studies can have wider applications at university level, with benefits in literary and cultural studies (Boulton, forthcoming; Boulton & Wilhelm, 2006).

APPENDICES

APPENDIX 1: Questionnaire on two adverbs: ALMOST AND NEARLY

We, at the CRAPEL, are currently working on a project for the Cercle des Linguistes Anglicistes de Nancy2. We should be extremely grateful if you accepted to answer this short questionnaire.
Thank you!

1) Are you a native speaker of English?

Yes☐ ☐ No

2) Instinctively, would you say there is a difference between 'Almost' and 'Nearly'?

Yes☐ ☐ No

If you answered 'Yes': What would the difference(s) be?

_____
_____
_____
_____
_____
_____

PTO

3) As an afterthought, would you say that:

a. one of them had a more negative connotation? Tick the boxes accordingly.

Yes Yes
Almost Nearly
☐ No ☐ No

b. one of them was used more frequently? Tick the boxes accordingly.

Yes Yes
Almost Nearly
☐ No ☐ No

c. one of them was more formal? Tick the boxes accordingly.

Yes Yes
Almost Nearly
☐ No ☐ No

d. one of them was more British than American? Tick the boxes accordingly.

Yes Yes
Almost Nearly

☐ No ☐ No

4) Do any of the following sentences seem wrong to you? If so, write “WRONG” after the sentence(s) that do not satisfy you.

a. The client will almost definitely ask you this
b. The client will nearly definitely ask you this

c. The new pool owner will almost certainly feel some concern for the water-lilies during the winter
d. The new pool owner will nearly certainly feel some concern for the water-lilies during the winter

e. He was nearly completely bald
f. He was almost completely bald

g. You're not nearly as clever as you think you are
h. You’re not almost as clever as you think you are

APPENDIX 2: First 30 concordances for *almost* and *nearly*, from VIEW (News).

ALMOST

an end. But in return for its rich density of lovingly dwelt-on,	**almost**	novelistic detail, what you lose is any sense of a strong
seasoned professional. Edmund Barham makes Cavaradossi	**almost**	as formidable, a convincing aristocrat, tender with Tosca, arrogant
crudely stereotyped hack and his tediously distracted wench	**almost**	became convincing. But one was left wondering at the decision to
--; that the heroine of poet Carol Rumens's first play	**almost**	Siberia twists apart in a central crisis scene. She is,
it has rather more inside it than the play does.	**almost**	Siberia is a glasnost spin-off piece: a genre that was bound
United, the second-largest US airline, its investment represents	**almost**	80 per cent of the equity in the deal. Under the apparent
at £282.2m. A rise of 14p amid speculative interest added	**almost**	£10m to the value of the bathroom fittings and showers business on
newsletter, the UK is still by far the largest market with	**almost**	700,000 subscribers split between Racal Telecom's Vodafone and
per 1,000 inhabitants, Norway leads the penetration stakes with	**almost**	40 handsets per 1,000 inhabitants. Apart from the UK and
to FFr1.34bn last year. Over the same period new business	**almost**	doubled to FFr68.4bn, total loans and other financings grew from
British cuppa and imparts that unique lorry-driver flavour, has	**almost**	doubled from 78p a kilogramme to 135p. Medium qualities have also
's annual meeting, adjourned until 10 October, will now	**almost**	inevitably be put back for a further four to six weeks in the light of
attack on the Prime Minister, and one which is improper	**almost**	to the point of being constitutional. They want the Prime Minister
in Wandsworth, the significance of which appears to have been	**almost**	totally ignored by the media. Electors are always prepared to criticise
Gorbachev era, negotiated rather than unilateral disarmament is	**almost**	always preferable. On the economy, there is still work to
However, at least 10 of my cultural heroes would appear on	**almost**	all lists, and perhaps 30 would appear on most. Nobody
anyone except soldiers and carefully vetted civilians --; it could	**almost**	be the Goddess of Democracy. Sculpted from white plaster like the
struggle, the cult of the model worker and even the	**almost**	evangelical rhetoric of the cold war are all back with a vengeance
Editor PAY RISES for managers are averaging 11 per cent,	**almost**	double the level of a year ago, and accelerating, according to
reached their "ceiling". Overall, salary rises have	**almost**	doubled in a year. The average increase reported in the March
years ago, less than 25 per cent had a car, now	**almost**	35 per cent. The survey analysed nearly 21,000 salaries of senior
The Satanic Verses, none of them put in context and	**almost**	all involving obscenities. The sheikh, a white-bearded man with the
wrangle. A hardline motion calling for black sections is	**almost**	certain to be thrown out on Friday and the party could be
officials say additional members attracted this year has	**almost**	offset that loss. In 1952-53 membership peaked at a million.
pub rock in a back room with adjoining outer bar for	**almost**	as long as the form has existed. Last Thursday night the pub
plethora of litigation surrounding Blues Brothers lookalikes	**almost**	matches that attached to the release of Wired, the unofficial
completely different species naturally, in the wild. There are	**almost**	certainly other surprises out there which, by definition, we can
and it was hard work, very painful." After	**almost**	two hours at the wheel yesterday, Senna limped away from his car
is cast for injured Richards By STEVE BALE ENGLAND will	**almost**	certainly be without their inspirational Lions No. 8, Dean Richards,
weeks. At Bridgend he looked in despair, and he	**almost**	admitted as much: "I hope we can pull it together

NEARLY

think, from the acoustic properties of the set. Johnson remains	**nearly**	ideal as the jealous, mercurial prima donna. Her powerful and
layer of pseudo-significance. THEATRE / Frozen attitudes:	**nearly**	Siberia --; Soho Poly By ALEX RENTON OME props ought to carry
property profits as earnings. We have made a bid of	**nearly**	£700m for a company with a book value of £200m --; we
grave. The Matanzimas were forced out after a military coup	**nearly**	two years ago. The benign ruler who took over, Major-General
Australian outback yesterday, bringing the total detained to	**nearly**	500 in three days of protests. Chemical Mace was used against
after the disaster, but Byelorussian activists say this was not	**nearly**	enough. Latecomers who continue to hope From ANNE
the Delta Commando and Col Jean-Marie Bastien-Thiry who	**nearly**	succeeded in blowing up General de Gaulle and his wife at Petit-

period from the mid-1970s it rose from just over £30,000 to	**nearly** £250,000, and analysts calculate that in 20 years - when all
so-called Next Steps --; for restructuring the Civil Service in	**nearly** 150 years. No successor could hope for such an earth-moving tenure
shortfall of contributions to contributory benefits, amounting to	**nearly** £500m a year. As for old colleagues resenting his intrusion on their
a car, now almost 35 per cent. The survey analysed	**nearly** 21,000 salaries of senior managers to supervisors at more than 900
Women Teachers reveals that under the new financing formula	**nearly** 60 per cent of secondary schools will be less well funded and about
authorities in England and Wales shows 882 schools will lose	**nearly** £55m between them annually --; an average of £62,663. In particular,
year showed 52 per cent favoured retention. The poll, of	**nearly** 2000 adults, showed 72 per cent favoured the retention of independent
of the financial problem is much smaller than in 1987 when	**nearly** 4,000 beds were closed, triggering the NHS review and an emergency
hurt in a fight with another male, Nikkie. For	**nearly** a week after the injury, whenever he was in Nikkie's
Richard Lee and a Jonathan Callard conversion, they	**nearly** allowed themselves to be knocked out of their stride by Neath's
summer. Environmental engineer Wheway slipped 2p to 142p.	**nearly** 95 per cent of its one-for-four rights issue was taken up,
which would put its losses so far at almost A$1.5bn.	**nearly** 15,000 tourism workers have lost their jobs and the industry has lost
home. He offers them to his friends. They all	**nearly** die. His mother says: "Why, oh why,
Bristol decided to test the theory. In a study of	**nearly** 500 women, they found that zinc supplements did not improve either
research suggests parental habits have altered little in	**nearly** 30 years. In 1958, some 62 per cent of mothers smacked
the age of seven, more than half the boys and	**nearly** a third of the girls were being hit at least once a week
palace shops, have had great success. Profits have increased	**nearly** eight-fold in that period to £500,000. The new chief executive,
. Grand Met was also the busiest option stock, claiming	**nearly** 2,800 contracts. It was suggested Goldsmith interests already had two
. They tried to kill her and her Cabinet, and	**nearly** succeeded, when they blew up the Grand Hotel, Brighton,
since almost all domestic broadcasts spill over borders	**nearly** all EC programming will be subject to the new
Peking, where he worked quietly and largely anonymously for	**nearly** a decade. But, as Mao grew increasingly suspicious of and
Mao called in the army to restore order. During his	**nearly** three years as effective day-to-day leader of the Cultural Revolution,
, the Communist party organ, published news of his death	**nearly** two weeks late but avoided any harsh commentary. Such signs of

REFERENCES

ALLAN, Rachel, 2006. "Data-driven learning and vocabulary: investigating the use of concordances with advanced learners of English". *Centre for language and communication studies, ocasional paper*, 66. Dublin: Trinity College Dublin.

ASTON, Guy, 1997. "Small and large corpora in language learning". In LEWANDOWSKA-TOMASZCZYK, Barbara; Patrick MELIA, *Practical applications in language corpora*. Lodz: Lodz University Press, p. 51-62.

BERNARDINI, Sylvia, 2001. "Spoilt for choice. A learner explores general language corpora". In ASTON, Guy, *Learning with corpora*. Houston: Athelstan, p. 220-249.

BIBER, Douglas; Susan CONRAD; Randi REPPEN, 1998. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

BONNET, Gérard ; Jacqueline LEVASSEUR, 2004. Evaluation des compétences en anglais des élèves de 15 ans à 16 ans dans sept pays européens. ftp://trf.education.gouv.fr/pub/edutel/dpd/noteeval/eva0401.pdf, last consulted 30/08/08.

BOULTON, Alex, 2007. "DDL is in the details… and in the big themes." In DAVIES, Mark; Paul RAYSON; Susan HUNSTON; Pernilla DANIELSSON, *Proceedings of the corpus linguistics conference: CL2007*. http://www.corpus.bham.ac.uk/corplingproceedings07/, last consulted 30/08/08.

BOULTON, Alex, 2009. "Testing the limits of data-driven learning: Language proficiency and training." *ReCALL*, 21/1.

BOULTON, Alex, in press. "Looking (for) empirical evidence for DDL at lower levels." In LEWANDOWSKA-TOMASZCZYK, Barbara, *Corpus linguistics, computer tools, and applications: state of the art*. Frankfurt: Peter Lang.

BOULTON, Alex, forthcoming. "Bringing corpora to the masses: Free and easy tools for language learning." In KÜBLER, Nathalie, *Selected papers from TaLC 2006*. Amsterdam: Rodopi.

BOULTON, Alex; Stephan WILHELM, 2006. "Habeant corpus – they *should* have the body. Tools learners have the right to use". *ASp*, 49-50, p. 155-170.

BREYER, Yvonne, 2006. "My Concordancer: tailor-made software for language learners and teachers". In BRAUN, Sabine; Kurt KOHN, K; Jobrato MUKHERJEE, *Corpus technology and language pedagogy: new resources, new tools, new methods. English corpus linguistics, 3*. Frankfurt: Peter Lang, p. 157-176.

British National Corpus. http://www.natcorp.ox.ac.uk/, last consulted 30/08/07.

BURNARD, Lou; Tony McENERY (eds.), 2000. *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang.

COBB, Tom, 2003. "Do corpus-based electronic dictionaries replace concordancers?" In MORRISON, Bruce; Christopher GREEN; Gary MOTTERAM, *Directions in CALL: experience, experiments, evaluation*. Hong Kong: Polytechnic University, p. 179-206. http://www.er.uqam.ca/nobel/r21270/cv/replace_conc.htm, last consulted 30/03/06.

FRANKENBERG-GARCIA, 2005. "Pedagogical uses of monolingual and parallel concordances". *ELT journal*, 59/3, p. 189-198.

GAST, Volker, 2006b. "The distribution of 'also' and 'too': A preliminary corpus study". In GAST, Volker. *The scope and limits of corpus linguistics: empiricism in the description and analysis of English. Zeitschrift für Anglistic und Amerikanistic*, 54/2. Würzburg: Königshausen & Neumann, p. 163-176.

GREMMO, Marie-José, 1995. Former les apprenants à apprendre. *Mélanges CRAPEL*, 22, p. 9-32.

HORNBY, Albert Sidney; Sally WEHMEIER, 2007. *Oxford advanced learner's dictionary*, 7th edition. Oxford: Oxford University Press.

HUNSTON, Susan, 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

JOHNS, Tim, 1991a. "Should you be persuaded: two examples of data-driven learning". In JOHNS, Tim; Philip KING, *English language research journal, 4: classroom concordancing*, p. 1-16.

JOHNS, Tim, 1991b. "From printout to handout: grammar and vocabulary teaching in the context of data-driven learning". In JOHNS, Tim; Philip KING, *English language research journal, 4: classroom concordancing*, p. 27-45.

JOHNS, Tim, 1997. "Kibbitzing one-to-ones (web notes)". *BALEAP: Academic Writing*, University of Reading, 29 November. http://www.eisu.bham.ac.uk/johnstf/pimnotes.htm, last consulted 30/03/06.

KJELLMER, Göran, 2003. "Synonymy and corpus work: On "almost" and "nearly"". *ICAME Journal*, 27, p. 19-27.

Language Log. http://itre.cis.upenn.edu/~myl/languagelog/, last consulted 30/08/07.

LEVY, Mike, 1990. "Concordances and their integration into a word-processing environment for language learners". *System*, 8/2, p. 177-188.

LOUW, Bill, 1997. "The role of corpora in critical literary appreciation". In WICHMANN, Anne; Steven FLIGELSTONE; Tony MCENERY; Gerry KNOWLES, *Teaching and language corpora*. Harlow: Addison Wesley Longman, p. 240-251.

McENERY, Tony; Richard XIAO; Yukio TONO, 2006. *Corpus-based language studies: an advanced resource book*. London: Routledge.

O'KEEFFE, Anne; Michael McCARTHY; Ronald CARTER, 2007. *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

OWEN, Charles, 1996. "Do concordances require to be consulted?" *ELT Journal*, 50/3, p. 219-224.

SINCLAIR, John (ed.), 2004a. *How to use corpora in language teaching*. Amsterdam: John Benjamins.

SINCLAIR, John, 2004b. "New evidence, new priorities, new attitudes". In SINCLAIR, John, *How to use corpora in language teaching*. Amsterdam: John Benjamins, p. 271-299.

THURSTUN, Jennifer; Christopher CANDLIN, 1997. *Exploring academic English: a workbook for student essay writing*. Sydney: CELTR.

TRIBBLE, Chris; Glynn JONES, 1997. *Concordances in the classroom* (2nd edition). Houston: Athelstan.

TSUI, Amy 2005. "ESL teachers' questions and corpus evidence". *International journal of corpus linguistics*, 10/3, p. 335-356.

Variation in English Words and Phrases. http://view.byu.edu/, last consulted 30/08/07.