Synonymy and corpus work: On *almost* and *nearly*

Göran Kjellmer
Göteborg University

Dictionaries are indispensable tools for a language learner. They will tell him what words mean and how they are distinguished semantically. What the dictionaries do not always do is specify how words are used idiomatically by native speakers, and in omitting to do so they can sometimes be misleading in that the information they offer may be necessary without being sufficient. Consider the case of *almost* and *nearly.*

Almost and *nearly* are close synonyms, as is clear from a few dictionary definitions:

almost all but; very nearly
nearly almost
(COD 1990)

almost very nearly but not completely
nearly almost, but not quite or not completely
(LDOCE 1995)

almost nearly but not quite
nearly almost or not completely
(CIDE 1995)

almost not quite; very nearly
nearly very close to; almost
(NODE 1998)

You use **almost** to indicate that something is not completely the case but is nearly the case. **Nearly** is used to indicate that something is not quite the case, or not completely the case.
(Cobuild 2001)

It appears from those definitions that the words are so synonymous that they are sometimes defined in terms of each other. If they are so close in meaning, one may well wonder if there are any differences between them in the way they are used. I will here discuss three aspects, viz:

- their frequency
- their style and text type preference
- their collocability

In order to do so I will make use of the CobuildDirect Corpus, where such aspects are convenient to deal with.

A. Frequency. *Almost* is more than twice (2.33 times) as frequent as *nearly* in CobuildDirect; it has 15,536 occurrences, while *nearly* has 6,666. Whether this difference in frequency between the words is significant or not (in a non-statistical sense) when it comes to their use is too early to say. But the difference could suggest that *nearly* is a more select and specialised word than *almost.*

B. Text type preference. CobuildDirect consists of twelve subcorpora, taken from different text types. The distribution of *almost* and *nearly* over those text types might shed some light on their characteristics of usage.

Table 1: Text type preference of *almost* and *nearly*

Almost:			
	usbooks	2256	401.0/million
	ukbooks	1939	362.1/million
	times	2027	351.7/million
	npr	1039	332.0/million
	oznews	1538	288.1/million
	ukmags	1337	272.7/million
	today	1422	270.9/million
	bbc	669	256.3/million
	sunnow	1374	235.9/million
	ukephem	536	171.6/million
	usephem	177	144.5/million
	ukspok	1222	131.8/million
Nearly:			
	npr	669	213.8/million
	bbc	532	203.8/million
	today	814	155.1/million

usbooks	794	141.1/million
times	758	131.5/million
sunnow	666	114.3/million
usephem	122	99.6/million
oznews	489	91.6/million
ukmags	449	91.6/million
ukbooks	461	86.1/million
ukspok	690	74.4/million
ukephem	222	71.1/million

The abbreviations for the sources are to be interpreted thus:

npr	= US National Public Radio broadcasts
today	= UK Today newspaper
times	= UK Times newspaper
usbooks	= US books; fiction & non-fiction
oznews	= Australian newspapers
bbc	= BBC World Service radio broadcasts
usephem	= US ephemera (leaflets, adverts, etc.)
ukmags	= UK magazines
sunnow	= UK Sun newspaper
ukspok	= UK transcribed informal speech
ukbooks	= UK books; fiction & non-fiction
ukephem	= UK ephemera (leaflets, adverts, etc.)

The middle column gives the raw frequencies of the words in the subcorpora, and the rightmost column converts those figures to comparable frequencies per one million words.

The table shows that a*lmost* prefers literary styles of writing (usbooks, ukbooks, times) and avoids more popular text types (sunnow, ukephem, usephem, ukspok), whereas *nearly* is more strongly favoured in the news media (npr, bbc, today). Neither of them is used much in spoken British English (ukspok).

C. Collocability. The corpus provides a facility by means of which the most significant collocates of any word in the corpus can be discovered. Four words on either side of a word are seen as its collocates. T-score calculations will then indicate their significance and sort the words accordingly. The most significant collocates of *almost* and *nearly* are given in Table 2.

Table 2: Collocates of *almost* and *nearly*

Collocates of *almost*			Collocates of *nearly*		
	n	**T-score**		**n**	**T-score**
certainly	650	24.637057	*years*	681	23.963591
as	1569	23.223282	*000*	354	17.239044
every	486	19.473519	*all*	533	16.802179
it	2164	19.356305	*million*	287	15.773051
years	562	18.235811	*half*	285	15.633906
impossible	269	16.002639	*two*	391	15.562322
was	1558	15.886701	*ago*	217	13.558705
an	800	15.398975	*three*	252	12.840255
all	707	13.871836	*every*	201	12.461522
immediately	208	13.724134	*hundred*	163	12.138026
certain	224	13.649910	*pound*	235	11.784722
half	275	13.621994	*thousand*	141	11.349662
000	305	13.507498	*dollar*	173	11.160020
entirely	149	11.767122	*cent*	157	11.070751
million	212	11.392851	*as*	534	11.056001
cent	199	11.085978	*per*	160	10.858025
ago	195	11.082241	*after*	240	10.485561
is	1509	10.987927	*four*	162	10.352990
always	211	10.592463	*year*	219	10.028319
dollar	210	10.275591	*for*	708	9.756185
any	303	10.048554	*months*	126	9.731231
exclusively	104	10.010189	*always*	128	9.146864
certainly	650	24.637057	*five*	131	9.061732
as	1569	23.223282	*six*	112	8.727735
every	486	19.473519	*30*	100	8.583709
it	2164	19.356305	*20*	100	8.492225
years	562	18.235811	*points*	86	8.350813
impossible	269	16.002639	*percent*	82	8.308806
was	1558	15.886701	*hours*	90	8.261594
an	800	15.398975	*40*	81	8.235720
all	707	13.871836	*billion*	75	8.178398
years	562	18.235811	*200*	71	7.956700
impossible	269	16.002639	*1*	136	7.912442
was	1558	15.886701	*died*	70	7.407585

an	800	15.398975	*50*	77	7.369675
all	707	13.871836	*300*	59	7.320331
immediately	208	13.724134	*2*	111	7.254769
certain	224	13.649910	*twenty*	67	7.203400
half	275	13.621994	*lost*	72	7.113234
000	305	13.507498	*quarter*	56	6.954470
entirely	149	11.767122	*weeks*	69	6.920944
million	212	11.392851	*killed*	59	6.845009
cent	199	11.085978			
ago	195	11.082241			
is	1509	10.987927			
always	211	10.592463			
dollar	210	10.275591			
any	303	10.048554			
exclusively	104	10.010189			
like	421	9.506938			
per	171	9.038828			
everything	127	8.795240			
seems	117	8.679364			
seemed	111	8.668655			
two	351	8.375584			
everyone	104	8.321448			
had	614	8.220298			
anything	134	8.190833			
percent	95	8.130223			
identical	67	7.996534			
completely	80	7.844077			
twice	78	7.771227			
could	288	7.512932			
daily	73	7.342104			
40	81	7.218745			
three	218	7.161602			
after	306	7.159486			
she	492	7.119327			
has	529	7.095944			

It appears from the table that typical collocates of *almost* and *nearly* are adverbs, adjectives, pronouns, prepositions, nouns and numerals in varying proportions. If we focus on the postcontexts of our words, Table 3 gives the distribution of the parts of speech immediately following them.

Table 3: Part-of-Speech-sorted words immediately following *almost* and *nearly*

	Adverbs	Adjectives	Pronouns	Verbs	Nouns	Numerals	Prep
Almost	2921	2743	561	2443	689/392	1840/2144	731
Nearly	395	366	106	1012	486/150	2633/2969	91
Almost *Nearly*	7.39	7.49	5.29	2.41	1.42/2.61	0.70/0.72	8.03

Almost is 2.33 times as frequent as *nearly.* If *almost* is much more, or much less frequent than 2.33 times the frequency of *nearly* in a given category, this would consequently be interesting.

Some preliminary conclusions are these. Adverbs, adjectives, pronouns and prepositions are typical postoccurring collocates of *almost*; nouns and numerals are typical postoccurring collocates of *nearly.* (The collocates preceding *almost* and *nearly* are most of the time, or at least very frequently, syntactically unrelated.)

However, some of these figures are misleading. Sums of money like $100 or £100 are given as "$ 100" and "£ 100" and classified as nouns. In this context it might be reasonable to regard them as numerals, and if they are treated as such, and if the difference in overall occurrence is taken into account, the difference between *almost* and *nearly* with regard to a following noun becomes uninteresting. (The adjusted figures are those following the slash.)

That the PoS categories are too general and conceal some important facts can be seen from the following experiment. There are 7.39 times as many adverbs following *almost* as there are following *nearly.* But what are those adverbs? Table 4 gives the beginnings of the lists "*almost* + adverb" and "*nearly* + adverb":

Table 4: Almost/nearly + adverb (RB)

a result-oriented judge who sided	almost always with the government against
matches twice a day, he would play	almost no more than an hour and a half and
if the US makes the final and plays	almost undoubtedly Australians in Saint
Algeria and Jordan met quietly,	almost obscurely in Rabat. All the

the Life of Andy Warhol,' set almost entirely to the music of the early
has laid out some scenarios which almost certainly would draw US military
did not foresee that this would lead almost immediately to a severe meat
phone call which was followed almost immediately by the bank calling to
by the PLO that the US would almost certainly have to veto. Yesterday
names, and one of those teams will almost surely be the national champion
markets. These people must rely almost entirely on state food stores for
he fell very fast. And people were almost deliberately staying away from
better you get, you know, and I'm almost as old as Sonny Boy Williamson was
The two countries went to war almost immediately, and when they finally
s happened. It's--it has happened almost everywhere. There's no parallel to
Lyman: Wirtz's mythical town is almost as densely populated as Lake
the road. One w--will be crushed, almost unidentifiably, and the other will
in with firm instructions and they almost intentionally, it seems to me
go in with firm instructions and almost intentionally, it seems to me
broadcast. When they are they're almost always used as propaganda.
a realist, but audiences and critics almost unanimously saw him as a romantic
Kuwait. Such a resolution would almost certainly be defeated, sources say.
non-entities, whose names were almost as long as the parts they play. In

in the wry music of Algeria and the nearly as famous as Oum Kalsoum but very
US Public Interest Research Group, nearly twice as many banks are charging
distribution rights, though not nearly as many as they had hoped for.
year, when Stemple earned nearly twice that amount. Other GM
makes furniture that could be put nearly anywhere. That's fortunate, since
hospitals. He found men were nearly twice as likely to undergo
black and white, and--and it's not nearly as polished as this <p> Simon: Mm-
immediate reaction would not be nearly as sympathetic and moreover that a
combined German team won't be nearly as good as the East German team was
I can.' By now her eyes burned nearly too high. I could see she was
fail ridiculously. But I don't feel nearly as foolish as the time I bought the
that six-year-old children were nearly as familiar with the Old Joe
as an insider or an outsider is not nearly as important as proving to voters
moves, which they say do not go nearly far enough, will be enough for
to the ancient Egyptians, he wasn't nearly as important as an earlier king,
Editor The parties are not nearly as powerful as they once were, but
US government is not spending nearly enough on industrial research and
student athletes graduate, a rate nearly twice the NCAA average. But
says the prosecutor did not go nearly far enough. She says all seven of
the kidnap-for-ransom gangs nearly always got what they asked for,

2nd OPERAGOER: Oh, nearly,	nearly, yes laughs <p> FRANKL: May
to a strict budget which is not	nearly as big as it was last year <p> At
HAZEL: Christmas is	nearly here and I'm exhausted already.

As we can see, *almost* is typically followed by manner adverbs (*obscurely, intentionally)*, time adverbs (*always, immediately*) and sentence adverbs (*undoubtedly, certainly*), whereas *nearly* typically occurs in the construction *not + nearly + as*, where *as* is classified as an adverb. There is thus a qualitative as well as a quantitative difference between *almost* and *nearly* in this respect.

Summary and conclusion. The differences found between *almost* and *nearly* are these:

- *almost* is much more frequent than *nearly* and is therefore likely to be less specialised than the latter;
- *almost* occurs more in literary styles of writing than in popular text types, whereas *nearly* is a preferred word in the news media. Neither of them is used much in the spoken language;
- their collocations distinguish the two words sharply, so that *almost* is characteristically followed by adverbs (*almost certainly*), adjectives (*almost impossible*), pronouns (*almost anything*) and prepositions (*almost by definition*), and *nearly* is equally characteristically followed by numerals (*nearly 200 people*).

These three aspects can be seen to be interrelated. *Nearly*, which occurs more often in the news media, where precision and factual information are more focused than in literary styles, is to some degree specialised in that it is preferably used to modify precise figures.

So 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. Even if they are closely related in meaning, corpus studies can show that they are used differently by idiomatic speakers of English. Identity or near-identity of dictionary definitions does not guarantee identity or near-identity of usage.

References

CIDE = Procter, Paul (ed.). 1995. *Cambridge International Dictionary of English.* Cambridge University Press.

Cobuild = Sinclair, John (editor-in-chief) 2001. *Collins COBUILD English Dictionary for Advanced Learners.* 3rd ed. Glasgow: HarperCollins.

COD = Allen, R.E. (ed.). 1990. *The Concise Oxford Dictionary of the English Language.* 8th ed. Oxford: Clarendon.

LDOCE = Summers, Della (ed.). 1995. *Longman Dictionary of Contemporary English.* 3rd ed. Harlow, Essex: Longman.

NODE = Pearsall, Judy (ed.). 1998. *The New Oxford Dictionary of English.* Oxford: Clarendon.