L.G. Mityushin

# Often or seldom?

Frequency of English expressions as an indicator of correctness.
A manual for learners of English who are not afraid of arithmetic.[1]

Contents



[1] This is a translation of the original Russian version (Л.Г. Митюшин. Часто или редко).

# 1. Introduction

This manual is about measuring the frequencies of English expressions in two large bodies of text. The first of these is the part of the British Internet accessible to the Yandex search engine; the second is over a million books in English digitized in the Google Books project.

By the frequency of an expression we mean the total number of its occurrences in texts. Why is this figure of interest to us? The reason is the close link between frequency and linguistic correctness: expressions of large frequency are almost certainly correct. We can put it differently: erroneous expressions are rare compared to the correct versions. In other words, in language "the majority is always right."

Of course, correct expressions can also be rare. The frequencies of several expressions with close meanings can differ greatly. For a non-native speaker, the most frequent expressions among those with similar meanings have two advantages: they are almost certainly correct, and they are likely to be stylistically neutral. Choosing the most frequent expression keeps to a minimum the probability of making a grammatical, lexical or stylistic error.

Besides measuring frequencies, the two collections of texts we deal with provide another useful tool. It is possible to search the texts for incomplete expressions containing one or more wildcards * ("stars"). In this case, the search finds expressions in which stars are replaced by some real words. Yandex shows the results as an array of quotations from the web pages with the expressions found; Google Books makes a list of all the expressions together with their frequencies.

Below we describe in detail how to work with Yandex (Chapter 2) and the Expression Database built for the Google Books text collection (Chapter 3).

## 2. Yandex

Suppose we want to estimate the frequency of an English expression[2] using the Yandex search engine. To do this, we open the web page www.yandex.com (interface language: English) or www.yandex.ru (interface language: Russian) and type our expression in quotation marks inside the search box. The quotation marks make Yandex look for pages where the words of the expression form a single group; without quotation marks, the words may occur far from each other on a page. Then, after one or more spaces, we type the word rhost:uk.* – it restricts the search to pages with the country code 'uk', that is those registered in the United Kingdom. This increases the probability that the text is written by a native speaker of English. For the sake of brevity, we will not write quotation marks and rhost:uk.* in examples of queries, but they are supposed to be always "invisibly present".

Then we click the button "Search" (or Russian "Найти") and get extracts from web pages that contain our expression. Besides that, its frequency is shown: the total number of pages where it occurs. The frequency is just what we are interested in. If the frequency is greater than 1000 it is rounded to thousands, and if greater than 1 000 000, to millions.

Sometimes the frequency number has a different meaning. If Yandex cannot find any pages containing the expression as a single group, it looks instead for pages where the words of the expression occur independently, and gives the warning: "Showing results for query without quotes." In this case, the frequency of the expression as a whole is zero.

Like other search engines, Yandex ignores punctuation and does not distinguish between small and capital letters. Thus we can carelessly type in the search box *english speaking countries*, and Yandex will find what is needed, that is, *English-speaking countries*.

---

[2] By "expression" we mean any sequence of words (possibly with punctuation marks); a single word may also be called "expression".

Queries can contain the wildcard character * ("star") surrounded by spaces, which represents one unknown word. For example, the query *english to * translation* gives pages with the expressions *English to French translation*, *English to Japanese translation*, *English to Zulu translation* and so on – anything you could imagine. A query can have two or more stars, then they should match the same number of real words on the page.

The following sections illustrate the use of simple queries and queries with stars.

## 2.1. Simple queries

Here we give examples in which two or more expressions are compared. The ratio of frequencies shows the difference in their degrees of usage.

The adjectives *high* and *tall* both mean "large in the vertical direction". It is well known that with a person, *tall* should be used. Indeed, for the expression *a tall man* Yandex finds 2000 pages; the number of pages for *a high man* is only 17, which is almost 120 times fewer. On the other hand, for example, the word *mountain* has opposite preferences: the page counts for *a high mountain* and *a tall mountain* are respectively 1000 and 89.[3]

Suppose we want to say that someone acted correctly using the expression *right thing to do*. Should there be an article, and if so, which one? We measure the frequencies of the three expressions: *it was the right thing to do* – 2000, *it was a right thing to do* – 8, *it was right thing to do* – 15. The first frequency is more than 100 times greater than each of the others.

We can compare expressions made up of the same words in different order. Word order in English is generally much stricter than in Russian, which may result in large differences in frequencies. For example, the frequencies of the expressions

[3] The frequencies in this version of the manual were measured on 7 July 2014.

*will always be able* and *will be always able* are 4000 and 27, with a ratio near 150.

What preposition should be used with the noun *exception*? Let us compare two expressions: *exception to the rule* and "more Russian" *exception from the rule*. Their frequencies are 7000 and 61, with a ratio over 100.

Here is another example of contrast between what is idiomatic in English and Russian. In English we often say *in the near future* while in Russian the corresponding idiomatic expression literally means *in the nearest future,* with the superlative form of the Russian adjective meaning *near*. The numbers of pages on the British Internet found by Yandex for *in the near future* and *in the nearest future* are respectively 140 000 and 841. On the other hand, the numbers of pages on the Russian Internet that Yandex finds for the Russian versions of these expressions are respectively 5000 and 4 000 000: the ratio is several times greater and in the opposite direction.

## 2.2. Queries with stars

The star is convenient when we don't know which words are typically used in a given context. We process a query with a star, then look through the quotations from the pages found by Yandex and see what real words occur instead of the star. Then we measure frequencies for queries in which the star is replaced by these words.

Our first example concerns the idiomatic use of prepositions. Which of them collocate with nouns like *company, factory* etc as a place where someone works or worked? We process a query with a star for the preposition: *worked * the company*. For this query Yandex finds 2000 pages, and we see that the quotations often contain *at, for, in* and *with*. Separate queries for the expressions with these prepositions give the following frequencies: *for* – 1000, *with* – 351, *at* – 222, *in* – 118. The query *worked * the factory* gives 420 pages; the quotations show that *factory* "loves" the prepositions *in* and *at* (their frequencies are 182 and 181).

Speaking of love and prepositions, let us demonstrate that simple-minded comparison of frequencies doesn't always give the right results. The noun *love* typically collocates with the prepositions *for* and *of*. *For* is usually used when we talk about feelings towards people, and *of* in other cases (*his love for his wife, his love of freedom*). The preposition *towards* may also occur instead of *for*, but it is much less frequent. What we cannot do is translate literally from Russian: *his love to his wife.* This is not correct English.

Let us try to support these facts with frequencies. To do this, we compare the expressions *his love for her* and *his love to her*. Their frequencies are 1000 and 124; the ratio is only 8. Why so little, if the combination *love* (noun) + *to* (preposition) really isn't used?

Everything falls into place when we look at the pages where *his love to her* occurs. It turns out that the prepositon *to* is almost always connected not with the word *love* but with some other word, usually a verb: *declares his love to her, is unable to express his love to her, in order to prove his love to her.* The conclusion is: when choosing expressions for comparison, we should make sure that they really represent the phenomena we are interested in. As the saying goes, "be careful what you wish for."

Returning to queries with stars, suppose we want to know which verbs collocate with the noun *potential* (= qualities that can be developed). We process the query *to * his potential* and get 1000 pages with these main "fillers" for the star in order of decreasing frequency: *fulfil, fulfill, realise, reach, achieve* (*fulfill* is the American spelling of *fulfil*). Their frequencies are 300, 173, 126, 99 and 45 respectively.

Which nouns are used in English to express the meaning "some irony", "a certain amount of irony"? We process the queries *a * of irony* and *an * of irony* (for words beginning with a vowel) and find that their frequencies are 3000 and 170. Quotations from the pages give these nouns (in order of decreasing frequency): *hint, sense, touch, trace, bit, twist, element* ... The most frequent expression *a hint of irony* has a frequency of 542.

To conclude this section, we find the adjectives typically used with the word *example*. Yandex gives a lot of results for the queries *a * example* and *an * example*: 359 000 and 70 000. This is the approximate list of the ten most frequent adjectives: *good, great, excellent, simple, classic, shining, superb, extreme, practical, nice*. The frequencies of the expressions decrease from 81 000 for *a good example* to 3000 for *a nice example*.

## 3. The Expression Database

In 2004 the Internet giant Google embarked on an ambitious programme of digitizing all the books ever published. By April 2013 over 30 million books had been scanned in the Google Books project.

The American linguist Mark Davies created a database of English expressions that occur in part of the Google Books data. He considered three types of books: published in the United States (155 billion words in 1.3 million books), published in Britain (34 billion words) and "one million books" (89 billion words; in this corpus the books were selected in such a way that their years of publication were more uniformly distributed in time). In these books, all expressions were found that contain up to 5 words and occur at least 40 times. Davies stored them in a database able to answer a wide range of questions about the information it contains.

We call this database ED (Expression Database). The examples below are taken from the British corpus; the American corpus and "one million books" can be dealt with in the same way.

At this point we advise the reader to open the page goglebooks.byu.edu and perform the actions described in the text. First we choose the corpus: British. An empty table appears, with years 1810, 1820, ... , 2000 written above the columns. The space under the table is occupied by the HELP area. In the left-hand part of the screen there are five fields, one below the other: DISPLAY, SEARCH STRING,

SECTIONS, SORTING AND LIMITS, and OPTIONS; their functions will be explained as the need arises. First we will try to work with ED in the same way as with Yandex.

### 3.1. Simple queries

Consider the queries from Section 2.1, in which all of the words are known. We take the expression *a tall man* and type it in the field SEARCH STRING. Quotation marks are not needed as ED always regards queries as connected groups of words. After the SEARCH button is pressed, a table appears in which the left-hand column has the heading WORD(S). This column contains three versions of our expression with different capitalisation: *a tall man*, *A tall man* and *a tall Man*. Then in the TOTAL column we see the total number of occurrences of each version: 17 230, 2135 and 61 respectively. Then the table shows how often the three versions of the expression appear in the books published in each decade from 1810–1819 to 2000–2009. The bottom row contains sums of numbers in the columns; for the column TOTAL it shows the overall number of occurrences of the expression: 19 426.

Now we try *a high man*. This time the table contains only one version of the expression: that with small letters; the total number of its occurrences in the books published in 1810–2009 is 109.

Note that the numbers in the ED table and the frequencies reported by Yandex have different meanings. ED counts the total number of occurrences of the expression in the texts while Yandex counts the number of pages where the expression occurs without taking into account how many times it appears on each page. As an expression may appear on a page more than once, the result produced by Yandex is lower than the total number of occurrences. However, in most cases the difference between these figures is no more than a few per cent, and we regard it as negligible. Hence we will use the same term "frequencies" for the numbers

produced by ED (strictly speaking, this is more accurate than in the case of Yandex).

Comparing the results for the expressions *a tall man* and *a high man* obtained in two different ways, we see that the ratio of their frequencies is 178 in ED and 116 in Yandex. We can say that ED shows a stronger preference for *a tall man* than Yandex.

The period 1810–2009 is set in ED by default. The user can choose a different period in the field SECTIONS. On the left, three standard intervals are shown: 1980s–2000s, 1800s–2000s (actually represents the years 1810–2009) and 1500s–2000s; using the mouse and the Shift key, any other sequence of decades can be chosen (not less than two: the system refuses to work with a single decade).

As we are interested in modern English, from now on we will work with the period 1980–2009; to do so we click on the first of the three standard periods. The corresponding volume of text in the British part of ED is 10.5 billion words. In this portion, *a tall man* occurs 3790 times and *a high man*, 15 times; the frequency ratio is 253, which is even greater than for the period 1810–2009.

A query can contain several options for a certain word divided by a vertical bar meaning "or". Thus, the queries *a tall man* and *a high man* can be merged into the single query *a tall|high man.* However, in this case expressions with different capitalisation are not considered, and for *a tall man* we get a smaller number: 3185.

Moving on to the example *it was the / a / Ø right thing to do* (*Ø* denotes "the empty word", that is, absence of a word), we find that ED cannot work with these expressions as they contain more than 5 words. To get around this problem, we consider the shortened expressions *the right thing to do* and *a right thing to do.* For them Yandex gives frequencies 20 000 and 149; ED gives 7234 and 54, with practically the same ratio. Note that we cannot do this with the expression *it was right thing to do* because the shortened expression *right thing to do* does not imply the absence of an article – on the contrary, it includes the longer expressions *the right thing to do* and *a right thing to do* as special cases.

For the other expressions from Section 2.1, ED gives these frequencies:

| | | |
|---|---|---|
| *exception to the rule* | 5193 | 5193/78 = 67 |
| *exception from the rule* | 78 | |
| *will always be able* | 1104 | 1104/8 = 138 |
| *will be always able* | 8 | |
| *in the near future* | 41 923 | 41923/138 = 304 |
| *in the nearest future* | 138 | |

The frequency ratios given to these pairs by Yandex are 115, 148 and 166.

Unlike Internet search engines, ED distinguishes between small and capital letters and pays attention to punctuation. Punctuation marks (including hyphens and apostrophes) are looked on as separate words and should be surrounded by spaces when typing a query. There is one exception: the combination "apostrophe + s" at the end of a word, which is typed as usual. Thus, the expression *he'll come* should be typed with spaces before and after the apostrophe, and *he's come* without spaces. An error is punished by producing an empty result, which means that the frequency of the expression in the corpus is less than 40.

Speaking of punctuation, we should note one problem with the original Google Books data: it does not include expressions with commas. Accordingly, in ED queries commas are not allowed (lead to empty results). This makes it difficult, for example, to count frequencies of parenthetical expressions, which are normally used with commas. So we cannot compare *Frankly, I think* and "the Russian" *Frankly speaking, I think* (the Yandex frequencies for these expressions are 3000 and 25).

## 3.2. Queries with stars

Generally, ED handles a simple query just like Yandex, that is, returns the number of occurrences of the expression in the texts. Queries with stars are dealt

with quite differently: ED finds expressions where the stars are replaced with real words and outputs them all (or a substantial part of them) together with their frequencies. By default the expressions are listed in order of decreasing frequency.

Let us see what ED does with the queries from Section 2.2. For *worked * the company* one click on the SEARCH button gives a list of 5 expressions with the prepositions *for, with, in, at, by*; their respective frequencies are 526, 102, 58, 44 and 3. The query *to * his potential* also gives 5 results: *fulfil, develop, realize, achieve, realise* with frequencies 91, 55, 54, 39, 36. For *a|an * of irony* the list is longer: 33 expressions. It begins with the "fillers" *touch, hint, trace, sense, note, degree, kind, element, twist, tinge,* their frequencies being 650, 342, 337 ... , and ends with *suggestion, smile, spice, look* having the frequencies 4, 3, 3, 1. The total frequency of the 33 expressions is 2736.

For the last query, *a|an * example*, the number of resuls exceeds 100, but only the first 100 are shown in the table. This default limit on the table size can be changed in the OPTIONS field. We press the button CLICK TO SEE OPTIONS, type 1000 in the # HITS field and press SEARCH again.

This time the table shows the whole list of 657 expressions; their total frequency is 270 567. At the top are the expressions with the words *good, typical, excellent, simple, classic, prime, fine, clear, early, perfect, extreme, interesting*; their frequencies are 55 088, 11 141, 11 022, ... .

Unlike in Yandex, where stars only replace whole words, in ED a star can replace any string of letters inside a word (including an empty string). Let us make an experiment: set the number of results (# HITS) to 100 000 and input the query *s** . We almost get what we wanted, namely all the words beginning with *s* in order of decreasing frequency, but only 4000 of them instead of the complete list. This is an absolute limit on the number of results shown by ED.

Let us process a more realistic query, for example, *answered *ly* . This time we have "only" 393 results; the most frequent "fillers" in this expression are *only, correctly, simply, affirmatively, immediately, quickly*. There may be more than one star in a word. For example, the query **work** gives such words as

*metalworking*, *workaholic* and *unworkable* (but the first word on the list is, of course, *work*).

Besides the star, ED has another wildcard character: the question mark. When surrounded by spaces it denotes itself, that is, the corresponding punctuation symbol. But in company with other characters it represents any single character. It can be combined with the star to represent any non-empty string of letters. For example, by adding a question mark to the above query (*?*work**) we get the same list of words except those beginning with *work* , because on the left of *work* there should be at least one additional character.

Queries with question marks can be used in solving crossword puzzles, when some letters in the answer word or phrase are already known. Suppose, for example, that we are looking for a word which means 'expert in breaking the law' and has the form - - - *m* - - - *l* - - - *s* - - . We run the query *???m???l???s?* and get a list of 16 words, where the second one is just what we need: *criminologist*.

### 3.3. Word classes

Look more carefully at the queries in Sections 2.2 and 3.2: *worked * the company*, *to * his potential*, *a|an * of irony*, *a|an * example*. Actually, we were interested not in any "gap fillers" for these expressions but only in words of specific parts of speech: prepositions in the first case, verbs in the second, nouns in the third, and adjectives in the fourth. In the last three examples, the function word before the star was added to the query just to make sure that the "filler" belongs to the required part of speech.

In ED queries it is allowed to specify a part of speech explicitly. There is a list of word classes containing, among other categories, all the traditional parts of speech, and these classes can be used in queries instead of stars. Consider the expression *a|an * example*. The new query contains no article and starts directly with the adjective: [*j**] *example* , where the code [*j**] means "any adjective". We

get a list of 1171 expressions with the total frequency of 606 311. Comparison of the old and new lists shows that at the top they contain almost the same adjectives, but their positions may differ by a few places up or down. However, there are also some significant differences: for example, the eighth place on the new list is occupied by the expression with the word *best*, while the old list contains no expressions with superlative adjectives (due to the indefinite article).

The main parts of speech have these codes: [n*] – noun, [v*] – verb, [j*] – adjective, [r*] – adverb, [p*] – pronoun, [i*] – preposition, [c*] – conjunction. There are also narrower categories: for example, [nn1*] denotes a single common noun, [vm*] a modal verb, and [jjr*] a comparative adjective. The list of word classes and their codes is given in the Appendix. Instead of typing a code on the keyboard, it is possible to open the drop-down list of word classes by clicking the POS LIST button in the SEARCH STRING field, and then click one of the classes: the result will be the same.

Codes of word classes may be combined with partly defined words using a dot as a separator. For example, the query *un*.[j*] *ness* gives *unfinished business*, *unhappy consciousness*, *underlying illness* ... – 1125 expressions in all, which occur 37 669 times.

From a technical point of view, ED deals with word clases in a rather complicated way. The problem is, the original corpus of expressions on which ED is based contains no information about word classes. ED gets the necessary data from a different source: the Corpus of Contemporary American English (COCA), and assigns a word a certain class if in COCA it is assigned this class in at least 50% of cases. As a consequence, search results contain some "noise". However, this is outweighed by the advantages associated with the use of word classes in queries.

### 3.4. Other features

In a query, we can use the dictionary form of a word (called "lemma") enclosed in square brackets. In this case ED looks for expressions with all the forms of the word in brackets. Thus, for the query [*go*] *to school* the expressions with the words *go, goes, went, gone, going* are produced. In the case of adjectives, the comparative and superlative forms are considered: the query [*high*] *prices* gives the expressions *high prices, higher prices* and *highest prices.*

If we put an equals sign before a word in square brackets, ED will look not for the forms of this word but for its synonyms. For example, the query [=*beautiful*] *woman* gives 14 expressions with the adjectives *beautiful, attractive, lovely, handsome, wonderful, charming, striking, delightful, gorgeous, magnificent, stunning, exquisite, superb, pleasing* (in order of decreasing frequency). For the query [=*good*] *weather* ED finds 26 expressions (*fine, good, fair, clear, sunny, mild*...), for [=*bad*] *weather*, 23 expressions (*bad, adverse, severe, poor, appalling, harsh*...).

For this type of query, ED only finds expressions in which the synonyms are in the dictionary form. So, the query [=*find*] *the answer* gives 5 expressions with the verbs *find, get, discover, obtain* and *understand* in the infinitive; other forms of these verbs are not considered. Neither should we expect that all the words that are synonymous in a specific situation will be found. For example, the adjectives *strong* and *thick* can be regarded as synonyms when combined with the noun *accent* (*thick German accent = strong German accent*), and the expressions *strong accent* and *thick accent* are both present in ED for the period 1980–2009 (with frequencies 562 and 235). However, the expression *thick accent* is not among the results for the query [=*strong*] *accent*. The fact is, when processing such queries ED makes use of generalized sets of synonyms, which have similar meanings in a wide range of situations and not just in combination with certain limited groups of words.

In front of a query element, a minus sign (or hyphen) can be put, which means

negation. For example, the query *answered -*ly* produces expressions in which *answered* is followed by any words except those ending in *ly*. The minus can be combined with any query elements except ordinary words: the query *answered -correctly* gives an empty result.

### 3.5. Miscellanea

ED is free and available to everyone. However, after the first 10–15 queries users are invited to register. It can be done either via the link we ask that you register or on the page corpus.byu.edu (which lists all the corpora created by Mark Davies) after pressing the button "Register" in the menu on the left.

The opportunities offered by ED are described in detail in the HELP area, which occupies the bottom right-hand part of the screen. In the previous sections we talked about what was necessary for our specific purposes. However, ED also has a wide range of functions and properties that could be of interest in a broader context. Below we describe two of them, and refer the reader to the HELP for the others.

The DISPLAY field provides two options: LIST and CHART. The default one is LIST, which means that the results are presented as a numerical table. The rows of the table correspond to the expressions found, the columns correspond to decades within the considered period of time, and each cell contains the number that shows how many times a given expression occurred in the books published in a given decade. There is also a column that shows the total number of occurrences of each expression over the whole period. It is in this mode that all our figures were obtained.

If we choose the other option, CHART, the results will be presented as a graph: a line of rectangular blocks where the height of each block is proportional to the relative frequency of the given expression in the books published in the given decade. By relative frequency we mean the absolute frequency of an expression

(the number of its occurrences in the texts) divided by the total number of words in the texts. It could be said that the relative frequency of an expression is equal to the probability of its appearance at a randomly chosen point in the text. This mode shows in graphic form how the use of the expression changed over time. It should be noted that for queries that produce two or more expressions, the graph shows the total frequency of them, which means that, unlike the table, the graph gives information about the query as a whole.

The cells of the table created in the LIST mode are coloured in different shades of blue. This is particularly noticeable when the period of time is long, such as 1810–2009. The intensity of colour in the cells of a row connected with a certain expression shows how much it was used in different decades. Dark blue patches speak of the time when the expresson "flourished", pale strips indicate its "decline".

The intensities are chosen as follows. For each expression, its relative frequencies in various decades are calculated, and the maximum relative frequency over the whole period is found. Then for each cell, the corresponding relative frequency is divided by the maxmum relative frequency. The results fall into one of five groups: 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8 and 0.8–1 (border cases are included in groups with greater values). Accordingly, each cell gets one of five intensities of blue, from very pale (0–0.2 range) to quite dark (0.8–1 range).

## 4. Conclusion

We have learned how to measure frequencies of expressions in two very large text corpora: the part of the British Internet that Yandex works with and the books covered by Mark Davies's Expression Database.

Although Yandex provides a narrower range of functions than ED, it also has some advantages. First, Yandex can work with expressions of any length. Second, it works very quickly: with a good Internet connection you get the answer

instantly; in the case of ED, processing a query with many results can take several seconds. Third, Yandex ignores punctuaton, which is often convenient (remember the problem that ED has with commas).

The most popular Internet search engine – Google – also has these properties. Why, then, did we choose Yandex? Unlike Google, it doesn't search the whole Internet, but deals mainly with its Russian part. The reason is simple: Google, remarkable as it is in many respects, is notoriously unreliable as a tool for measuring frequencies. Besides that, Google works with the star in a different way: it allows the star to replace more than one word, and as a consequence, produces many irrelevant results.

The part of the British Internet visible to Yandex contains, very roughly, 20–30 billion words. This is dozens of times greater than the total number of words that people hear or read throughout their lives (and for a person, this stream of text is the main source of information about the mechanics of the language). A corpus of this size would contain numerous examples of usage which serve as models for those who learn English as a first language. Of course, it would be preferable to have statistics for the whole British Internet; however, the present situation should also be considered satisfactory.

The Internet is a dynamic and constantly changing ocean of information. The frequencies of the same expressions can change noticeably even over a short period of time. The important point is that the qualitative differences between frequencies remain the same: if expression A was much more frequent than expression B, it will almost certainly continue to be so. Note that from this point of view ED is an ideal system: its information is static, and all measurements are perfectly reproducible.

Besides access to large text corpora, the Internet Era provides language learners with other opportunities. For example, today hundreds of radio stations in English-speaking countries broadcast their programmes over the Internet. The most useful for learners are "conversational" stations with a limited amount of music, such as BBC Radio 4 (www.bbc.co.uk/radio4) or BBC London

(www.bbc.co.uk/bbclondon).

The Internet gives free access to online versions of the well-known English-English dictionaries for learners: Cambridge (dictionary.cambridge.org), Longman (www.ldoceonline.com), Macmillan (www.macmillandictionary.com), Oxford (www.oxfordlearnersdictionaries.com) (the dictionaries are listed in alphabetical order). They contain the same information as their printed counterparts but are more convenient to use. There is also a new function: pronunciation of words is not only shown in phonetic notation but can be heard as well.

These dictionaries give very good definitions of words and illustrate them with numerous phrases and sentences. However, some types of words are by their nature difficult to describe in dictionaries – for example, names of plants and animals. Their definitions contain only very general information which is not sufficient to understand what these objects really are. Here again, the Internet can be helpful: we can search it for visual images associated with words. We take Google or some other search engine, type the word we are interested in – for example, the name of the bird *kingfisher*, – then press the button "Images" and get a lot of pictures of this interesting bird. Another example is colour names: the difference between *"brown hair"* and *"auburn hair"* becomes quite clear when we run these queries on Google and get hundreds of photographs of women with these colours of hair.

To sum up, we can say that access to very large text corpora opens new possibilities for learners of English. A good command of language requires not only the ability to find words that have the desired meaning, it is also necessary to be able to use words in idiomatic combinations. The methods and techniques described in the manual may help to achieve this goal.

If you have found this manual useful,
please recommend it to your friends and colleagues.

Appendix

Word classes and their codes in queries.

The first column contains word classes as they are denoted in the drop-down POS LIST (= Part of Speech List); the second shows their codes used in queries.

| | | |
|---|---|---|
| noun.ALL | [nn*] | common noun |
| verb.ALL | [v*] | verb |
| adj.ALL | [j*] | adjective |
| adv.ALL | [r*] | adverb |
| neg.ALL | [xx*] | negative particle (*not*) |
| art.ALL | [at*] | article |
| det.ALL | [d*] | determiner, including pre- and post-determiners (*all, many, much, few, little, some, several, any, another, both, each, every, same, this, that, these, those, what, whatever, whatsoever, which, whichever, whichsoever, whose*) |
| pron.ALL | [p*] | pronoun, except possessive adjectives |
| poss.ALL | [app*] | possessive adjective (*my, our, your, his, her, its, their, thy*) |
| prep.ALL | [i*] | preposition |
| conj.ALL | [c*] | conjunction |
| noun.ALL+ | [n*] | noun |
| noun.SG | [nn1*] | singular common noun |
| noun.PL | [nn2*] | plural common noun |
| noun.+PROP | [np*] | proper noun |
| verb.BASE | [vv0*] | dictionary form of a lexical verb |
| verb.INF | [v?i*] | verb in the infinitive |
| verb.INF/LEX | [vvi*] | lexical verb in the infinitive |
| verb.MODAL | [vm*] | modal verb |

| | | |
|---|---|---|
| verb.3SG | [v?z*] | verb in the 3rd person singular |
| verb.ED | [v?d*] | verb in the past |
| verb.EN | [v?n*] | past participle |
| verb.ING | [v?g*] | present participle |
| verb.LEX | [vv*] | lexical verb |
| verb.BE | [vb*] | verb *be* |
| verb.DO | [vd*] | verb *do* |
| verb.HAVE | [vh*] | verb *have* |
| adj.CMP | [jjr*] | comparative adjective |
| adj.SPRL | [jjt*] | superlative adjective |
| adv.PRTCL | [rp*] | adverbial particle |
| adv.WH | [rrq*] | interrogative/relative adverb (*how, why, where, when* ...) |
| pron.INDF | [pn1*] | indefinite or negative pronoun (*everything, anything, something, nothing, everybody, anybody, somebody, nobody, everyone, anyone, someone, nil, nought*) |
| pron.PERS | [pp*] | personal pronoun (*I, we, you, he, she, it, they, thou, ye, me, us, him, her, them, thee, mine, ours, yours, his, hers, theirs, myself, ourselves, yourself, yourselves, himself, herself, itself, themselves, thyself*) |
| pron.WH | [pnq*] | pronoun *who* and its derivatives (*who, whom, whoever, whosoever*) |
| pron.REFL | [ppx*] | reflexive pronoun (*myself, ourselves, yourself, yourselves, himself, herself, itself, themselves, thyself*) |
| num.CARD | [mc*] | cardinal number (*one, two, three, four, ... ; 1, 2, 3, 4, ... ; I, II, III, IV, ...* ) |

| | | |
|---|---|---|
| num.ORD | [md*] | ordinal number (*first, second, third, fourth, ... ; 1st, 2nd, 3rd, 4th, ...* ) |
| conj.CRD | [cc*] | coordinating conjunction |
| conj.SUB | [cs*] | subordinating conjunction |
| interj | [uh*] | interjecton |
| PUNC | [y*] | punctuation mark |