ISSN 0320–2372. ІНОЗЕМНА ФІЛОЛОГІЯ. 2018. Випуск 131. С. 36–45
INOZEMNA PHILOLOGIA. 2018. Issue 131. P. 36–45

УДК 811.111'366.58-13 DOI: http://dx.doi.org/10.30970/fpl.2018.131.2137

LINGUISTIC VARIATION OF ENGLISH COGNITIVE VERBS

Iryna Dilai

Ivan Franko National University of Lviv,
1, Universytetska St., Lviv, Ukraine, 79000,
irynadilay@gmail.com

The objective of this paper is to follow the linguistic change of English cognitive verbs overtime, their variation across dialects, genres, registers, as well as gender related usage preferences. The data for the study of the linguistic variation has been taken from the BYU corpora (GloWbE, COCA, COHA, BNC, CORE) and MICASE. Leveraging both corpus-based and corpus-driven approaches enabled addressing the problem comprehensively and identifying salient (socio)linguistic variables within the class. The distribution of the top English cognitive verbs across dialects, genres, registers and their overtime change suggest that linguistic variation is highly typical of the verbs under analysis and follows distinctive patterns requiring deeper scrutiny. The tendency towards more abstract senses, deviation from grammatical norms and domain-specific usage uncovers hidden semantic processes triggered by the external factors. The findings can benefit the study of linguistic variation in general and be used by lexicographers and language instructors.

Keywords: linguistic variation, language change, language variety, linguistic variable, corpus.

Introduction. Modern dictionaries and thesauri oftentimes fail to reflect the dynamics of the language vocabulary, its genre and register stratification, overtime language changes, vocabulary preferences and gradual demise of certain senses and forms. Likewise, they tend to disregard language variation and crucial sociolinguistic aspects of vocabulary functioning. Thus, linguistic variation as "the main cause of linguistic change" [1, p. 47] needs constant attention on the part of lexicographers, language instructors and sociolinguists in particular. Cognitive verbs as a focus of this study constitute an indispensable part of the vocabulary of any language. Having arisen from concrete vocabulary, they underwent gradual semantic change towards more abstract senses. Their application in knowledge-based terminology, academic discourse, Bloom's classification, philosophy, psychology, AI and other domains exhibit the tendency towards more sophisticated and domain-specific senses. However, so far, there seems to be no explicit research into the linguistic variation of English cognitive verbs based on corpus data, which can reveal their specific functioning.

Theoretical background. The concept of a rigid linguistic norm appears to be unfeasible. According to W. Labov, variation needs recognition in modern linguistics and linguistic change is an unavoidable phenomenon [8]. A number of research methods in language variation and change have been developed in modern linguistics based on collecting, analyzing and evaluating empirical data [7; 11]. In order to enhance validity of the findings, variation studies require solid quantitative support. Corpus linguistics reliance on natural language data was against the grain of the prevailing Chomskyan orientation. A surge of interest in computing

and remarkable achievements in its application in language studies made it plausible to collect massive databases of tagged texts in the form of corpora. Attested modern corpora can serve as valuable repositories of objective linguistic information, containing samples of both oral and written speech from various sources (COCA [14], BNC [13], MICASE [18], GloWbe [17]), diachronic information (COCA [14], COHA [15]), genre and register stratification (COCA, BNC, COHA, MICASE), dialects distribution (GloWbE), and other sociolinguistic parameters (gender in MICASE). Thus, following Hoffland and Johansson (1982), the corpus-based quantitative study of linguistic variation can "shed light on the utilization of words in different varieties of English, and can, in addition, serve as a starting point for stylistic and grammatical studies as well as for cultural observations" [5, p. 39].

Hence, the objective of this study is to follow the development of the cognitive verbs both overtime (20th–21st century) and within language varieties, speech genres and registers based on corpus data. The study stems from the assumption that implementation of corpus data for sociolinguistic purposes can bring to fruition the research into linguistic variation and language change as exemplified by the class of English cognitive verbs. Apart from tracing the general frequency preferences and the development of new senses dictated by the new epoch, it can uncover hidden sociocultural peculiarities underlying the use of the verbs under scrutiny.

Methodology. Dealing with vast text data requires explicit methodology. It presupposes the choice of the direction of the analysis. The deductive corpus-based approach relying on the preconceived hypothesis about language functioning is aimed at testing it on vast empirical evidence. Conversely, the inductive corpus-driven approach rests on the generation of a new theory based on the data analysed and tends to lead to unexpected findings. The combination of the former and the latter appears to be the most promising for this research. Moreover, applying statistical procedures based on corpora adds validity to the findings.

The key to the study of linguistic variation is the identification of linguistic variables, the most salient features arising from the differences between the compared corpora (or sections of a corpus). The examination of differences is predominantly frequency-based. However, as it was observed by Oakes, "in order to carry out comparisons, it is important to use corpora that are matched in as many ways as possible to reduce the number of independent variables that could impact on variation" [9, p. 160]. The chi-square test or using the normalization per million in BYU corpora can well serve this purpose.

Quantitative data. The time span, limited as it may seem (20th-21st century), encompasses crucial socio-political changes, duly reflected in the vocabulary. The epoch of Knowledge set in for good. What has changed in the perception of knowledge, mastering cognitive vocabulary, and within the vocabulary itself?

Previous research has identified over 600 English cognitive verbs [3] that have at least one "cognitive sense". The corpus-based methodology enabled identifying the most frequently used ones. The analysis of GloWbE revealed that 15 verbs with the highest frequency ranks are as follows: *know* – 3,077,622, *think* – 2,824,266, *believe* – 912,792, *understand* – 651,745, *decide* – 482,309, *remember* – 464,851, *feel* – 375,280, *realize* – 343,984, *study* – 311,735, *forget* – 247,900, *recognize* – 232,866, *guess* – 208,677, *imagine* – 196,364, *assume* – 192,760, *learn* – 172,627. Some of them, such as *believe, guess*, and *think* are synonymously related.

The most frequently used cognitive verbs are assumed to dictate the tendencies underlying the whole system of the cognitive verbs. The diachronic study based on COCA revealed that the verbs *know, understand, remember, feel, imagine* and, particularly, *learn* and *study* show an increase in frequency over the last years. Simultaneously, the verbs *think, decide* and *forget* exhibit the reverse tendency. Similar results are mirrored in COHA. A closer qualitative examination of the language change requires a look at the language varieties.

Results and Discussion. The corpus-based approach to the study of the cognitive verbs consists of several stages. The first stage is the identification of the linguistic variables and their variants in the English cognitive vocabulary and a corpus-based study of these variables. A "linguistic variable" is a particular feature differently distributed geographically [12], it is "a structural unit that includes a set of fluctuating variants showing meaningful co-variation with an independent set of variables"[13, p. 334]. Previous research has pinpointed certain orthographic and phonetic variants across the language varieties, such as the use of the verb suffixes –ise and –ize. Still, it has not taken into consideration distinct semantic classes of words and focused mainly on the principal varieties, such as British English and American English. Since English as a global language has developed numerous varieties, which in turn are highly dynamic meeting the needs of the society, any research entails the analysis of the underlying sociolinguistic processes.

The second stage of the study implements a corpus-driven approach. Without having any specific preconceived theories about the functioning of the cognitive verbs across varieties, the overall distribution of the verbs was searched in the corpus and thus certain variables have been identified.

The data for the study of linguistic variation has been taken from GloWbE, the Corpus of Global Web-based English, released in 2012–2013 by Mark Davies, BYU, which contains 1.9 billion words from 20 countries where English is widely spoken [17], thus includes the following varieties: English spoken in the USA (US), Great Britain (GB), Ireland (IE), Australia (AU), New Zealand (NZ), India (IN), Sri Lanka (LK), Pakistan (PK), Bangladesh (BD), Singapore (SG), Malaysia (MY), the Philippines (PH), Hong Kong (HK), South Africa (ZA), Nigeria (NG), Ghana (GH), Kenya (KE), Tanzania (TZ) and Jamaica (JM).

The corpus-based analysis has established frequency counts across the varieties for the following linguistic variables:

- *-ise/-ize* verb suffixes;

  The *-ize* verbs are preferred in American English, Hong Kong and Bangladesh, whereas *-ise* forms are proliferated in British English, the Irish variety, New Zealand, India, South Africa and Kenya. However, it turns out that the other form also coexists in all the varieties. Besides, certain lacunas have been revealed, i.e., the absence of the correspondent

  -ise forms which are still found in dictionaries or reported elsewhere: *cognise, alphabetise, concretise, anonymise, antropomorphise, intellectualise, externalise* (though *internalise* is well-recorded), *memorise, universalize*, etc.

- *I guess* is mostly used in American and Canadian Englishes, as well as in the Singapore variety.
- The progressive usage of state verbs (*I am knowing, I am believing, I am understanding*) is spread in pidgin varieties, non-standard British and American English.
- Deviations of *I knows, he know, he knowed* type are mostly found in pidgin varieties.
- The absence of some verb forms: *cognize* is the only form of a respective verb found (no past or gerundial forms), which testifies to the lemma instability (associating senses with specific word forms [10, p. 6].
- Dialect words*:* Scottish *ken* (meaning 'know') in the British English section has been recorded 8 times. British dialects are not further classified in this corpus.

The corpus-driven approach revealed subtle instances of subliminal language patterns requiring further study and clarification. Among others, it discovered that some verbs are more spread in certain varieties, which can be attributed to their significance in the regions. For instance*, sensitize(-ise*) is predominantly found in Africa (Ghana, Tanzania, Kenya) and Jamaica. Perhaps, these countries, more than others, need to be made aware of the possible threats. For instance, "[…] regular campaigns can be held to sensitize the youth of the Area on family planning issues and also on sexually transmitted diseases" (GloWbE; GH G) [17]. *Idealize(-ise)* is infrequent in all varieties, except for the American one, which is consistent with the American cultural stereotypes. *Standardize(-ise)* yields comparatively significant results for the Hong Kong variety, which suggests the pressing need of the society. *Universalize* is hardly reported in New Zealand, Bangladesh, Ghana, Nigeria, Malaysia, and the Philippines.

The most frequent cognitive verbs are more frequently used in some varieties than in others. Table 1 features the distribution of *know* across the varieties. The figures in the Nigerian (NG) and American English (US) varieties stand out. However, the low frequencies in Sri Lanka (LK), Bangladesh (BD), Hong-Kong (HK) and Tanzania (TZ) are also linguistically relevant.

In order to account for the statistical distribution of the data the chi-squared value has been calculated. The contingency table for *know* has been built. The results are aligned with the normalization per million in GloWbE. The largest contribution to the overall chi-square made by each cell of the contingency table was made by the US section (23,458.4), followed by HK (4,056.3), NG (3,413.2), IN (2,572.1), and LK (2,555.9).

The distribution of *think* has reflected a similar tendency to that of *know* with the prevailing high frequency results for the US and noticeably low frequency for TZ (See Table 2).

SECTION	FREQ	PER MIL	
ALL	3077622	1,633.06	
US	753024	1,946.76	
CA	214497	1,591.63	
GB	622175	1,605.14	
IE	149328	1,478.07	
AU	243370	1,642.08	
NZ	121207	1,489.20	
IN	137269	1,423.50	
LK	62090	1,332.89	
PK	75692	1,473.55	
BD	51523	1,304.86	
SG	72454	1,685.97	
MY	70083	1,682.92	
PH	78759	1,821.08	
HK	49656	1,227.58	
ZA	66918	1,475.12	
NG	85018	1,993.57	
GH	55546	1,432.77	
KE	59812	1,456.65	
TZ	45022	1,280.49	
JM	64179	1,621.92	

Table 1. Distribution of *know* across GloWbE varieties

ISSN 0320–2372. ІНОЗЕМНА ФІЛОЛОГІЯ. 2018. Випуск 131

SECTION	FREQ	PER MIL	
ALL	2824266	1,498.63	
US	745325	1,926.85	
CA	190665	1,414.79	
GB	684665	1,766.35	
IE	131165	1,298.29	
AU	247302	1,668.61	
NZ	116530	1,431.74	
IN	103226	1,070.47	
LK	44359	952.25	
PK	50701	987.03	
BD	36026	912.38	
SG	68018	1,582.75	
MY	55212	1,325.82	
PH	54165	1,252.42	
HK	42854	1,059.42	
ZA	53744	1,184.71	
NG	53950	1,265.06	
GH	37256	960.99	
KE	39506	962.12	
TZ	28554	812.12	
JM	41043	1,037.23	

Table 2. Distribution of *think* across GloWbE varieties

Some other findings appear to be revealing. *Understand* has a more even distribution across varieties, though the US section outstands the others. HK variety has the highest frequency of *remember*, whereas NG has the highest frequency of *forget.* The US figures for *decide* are comparatively low, with the highest in favour of Singapore (SG). When it comes to imagine, the US frequency bar soars. When it comes to *feel*, SG and GB take the lead. Oddly enough*, learn* is used the most frequently in PH and HK, almost identically in Canada and in Tanzania but less in the US than in Canada. Study prevails in HK and TZ. Such preferences appear to be of sociocultural significance.

One more facet of the varieties' study is the analysis of synonyms. Thus, *believe* shows prevalence in NG and the US. The analysis of its synonyms revealed that *guess* is significantly more used in the US and SG. *Assume* is used mostly in the US, much more often than in Great Britain (GB), but *suppose* is used more in GB and Indian English (IE) than *assume*. *Reckon* shows the preference for AU and is scarcely represented in CA and US. Such kind of information appears indispensable for non-native speakers and language learners.

Genre and register study. According to Halliday (1989), there are two main types of variation in language, social and functional [4]. Dialects are characterized by social or regional variation, whereas register pertains to functional variation.

Register refers to specific lexical and grammatical choices as made by speakers depending on the situational context, the participants of a conversation and the function of the language in the discourse [4, p. 44].

Linguistic features, which are part of one speaker's dialect, might belong to a specific register for another speaker. Biber and Conrad [2, p. 4] define register as "situationally defined varieties" and distinguish four major registers: conversation, fiction, newspaper language, and academic prose, admitting, though, that registers can be defined at almost any level of generality. Biber's classification of registers is reflected in COCA, which contains the sections of spoken, fiction, newspaper, magazine and academic American English. Genres include both literary and non-literary text varieties, for example, short stories, novels, sonnets, informational reports, proposals, etc., and can be identified in corpora when looking at the data sources.

On analysing register-determined preferences of the cognitive verbs, one can observe that the verbs *know* and *think* are the most frequently used in the spoken register and the least in the academic section. It can be explained by different modality of the genres compared. At the same time, the verb *understand* almost equally prevails both in the academic discourse and spoken English. *Remember* and *forget*, as well as *feel* and *imagine,* are predominantly used in fiction, which reflects its narrative nature and the contribution of the verbs to building the narrative. *Learn* and *study* explicably show preferences for the academic discourse. A more fine-grained study of the genre distributions and the analysis of other verbs revealing subtle peculiarities of their functioning require a separate study.

CORE corpus (Corpus of Online Registers of English) by D. Biber, M. Davies and J. Egbert enabled following a more subtle register variation [16]. The corpus features over 30 registers amounting to over fifty million words. Though the compilers acknowledge the hybrid nature of the majority of online registers, it is still plausible to use this corpus to study variation across registers. The findings, however, are beyond the scope of this paper.

Gender study. The gender variation of the cognitive verbs was studied based on MICASE as it reflects this distinction. Table 3 shows the overall results for the top frequency cognitive verbs in a descending order.

verb	total	male	female
know	11,548	5,426	6,122
think	6,785	2,893	3,892
remember	889	443	446
understand	635	295	340
feel	598	210	388
study	436	204	232
learn	220	87	133
imagine	214	132	82
decide	138	54	84
forget	118	64	54
Total		9,808	11,773

Table 3. Gender-affected use of the top cognitive verbs

The verbs are ranked according to their total raw frequency in MICASE. Though women '*know, think, remember, understand, feel, study, learn'* and even '*decide*' more (at least by using these words) and men only '*forget*' more and '*imagine*' more, the total frequency ranking almost coincides with the male ranking. Females use *feel* more often than *understand* contrary to males, and *decide* more than *imagine*. The data is from the academic corpus of spoken English, so it might be the women's desire to catch up or even surpass men in the academic environment that makes them use more cognitive verbs than men. Since only one specialized corpus was used, the findings can be skewed and inconclusive. Thus, further research into gender-annotated corpora is needed in order to validate them.

Conclusions. The corpus study of linguistic variation heavily relies on frequencies and can hardly account for the use of every single word. Nonetheless, deeper research in this direction appears to be promising. The challenges of the variation study are dictated by the vast corpus data and the need to balance the corpora and normalize the findings. Besides, as it was observed by Oakes [9, p. 159], "certain types of stylistic research are not amenable to computer analysis, either because they consider linguistic features which involve a good deal of expert intuition, or because they consider linguistic features which are found only rarely". Thus, the correct interpretation of the findings requires solid (socio)linguistic background. Nevertheless, the implications of the corpus-based and corpus-driven variation study of English cognitive verbs are that its findings can be applied in teaching and learning ESL/EFL, translation studies, speech recognition and language processing systems, overview of dictionary entries, dialect studies, as well as benefit social sciences.

References

1. Biber D. University Language: a Corpus-based Study of Spoken and Written Registers / D. Biber. – Amsterdam, Netherlands : John Benjamins, 2006. – 261 p.
2. Biber D., Conrad S. Register, Genre, and Style / D. Biber and S. Conrad. – Cambridge : Cambridge University, 2009. – 344 p.
3. Dilay I. Cognitive verbs in English: a corpus-based Study / I. Dilay // What's in a Text? Inquiries into the Textual Cornucopia / ed. by A. Glaz, H. Kowalewski and A. Weremczuk. – Newcastle : Cambridge Scholars Publishing, 2012. – P. 217–230.
4. Halliday M. Spoken and Written Language / M. Halliday. – Oxford : Oxford University Press, 1989. – 128 p.
5. Hoffland K. Word Frequencies in British and American English / K. Hoffland and S. Johansson. – Bergen : Norwegian Computing Centre for the Humanities, 1982. – 234 p.
6. Hunston S. Lexis, wordform and complementation pattern: a corpus study / S. Hunston // Functions of Language. – no 10. – 2003. – P. 31–60.
7. Krug M. Research Methods in Language Variation and Change / M. Krug, E. Schlüter. – Cambridge : Cambridge University Press, 2013. – 536 p.
8. Labov W. Principles of Linguistic Change / W. Labov. – Volume II: Social Factors. – Oxford : Blackwell, 2001. – 592 p.
9. Oakes M. P. Corpus linguistics and language variation / M. P. Oakes // Contemporary Corpus Linguistics / ed. by P. Baker. – London, New York : Continuum International Publishing Group, 2012. – P. 157–183.
10. Sinclair J. Trust the Text: Language, Corpus and Discourse / J. Sinclair, R. Carter. – London : Routledge, 2004. – 224 p.
11. Using Corpora to Explore Linguistic Variation. Studies in Corpus Linguistics / ed. by R Reppen, S. M. Fitzmaurice, D. Biber. – Amsterdam and Philadelphia : John Benjamins Publishing Company, 2002. – xii +274 p.
12. Wardhaugh R. An Introduction to Sociolinguistics [7th ed.] / R. Wardhaugh. – Chichester, UK : Wiley-Blackwell, 2010. – 464 p.
13. Wolfram W. Variation and Language, an Overview / W. Wolfram // Encyclopedia of Language and Linguistics [2nd ed.] / ed. by E. K. Brown, A. Anderson. – Boston : Elsevier, 2006 – P. 333–341.

Sources

14. Davies M., 2004-. BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.
15. Davies M., 2008-. The Corpus of Contemporary American English: 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.
16. Davies M., 2010-. The Corpus of Historical American English: 400 million words, 1810–2009. Available online at http://corpus.byu.edu/coha/.
17. Davies M., 2013. CORE: Corpus of Online Registers of English. Available online at http://corpus.byu.edu/core/.
18. Davies M., 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. Available online at http://corpus.byu.edu/glowbe/.
19. Simpson R. C., Briggs S. L., Ovens J. and Swales J. M. The Michigan Corpus of Academic Spoken English. – Ann Arbor, MI : The Regents of the University of Michigan, 2002. Available online at http://quod.lib.umich.edu/m/micase/

Стаття надійшла до редколегії 06.09.2018
Прийнята до друку 25.11.2018

ISSN 0320–2372. ІНОЗЕМНА ФІЛОЛОГІЯ. 2018. Випуск 131

МОВНА ВАРІАТИВНІСТЬ АНГЛІЙСЬКИХ КОГНІТИВНИХ ДІЄСЛІВ

Ірина Ділай

Львівський національний університет імені Івана Франка,
вул. Університетська, 1, Львів, Україна, 79000,
irynadilay@gmail.com

Проаналізовано даних мовну зміну англійських когнітивних дієслів з плином часу, а також їхню варіативність залежно від діалекту, жанру, регістру та гендерних особливостей вживання. З'ясовано, що застосування корпусної методології дало змогу цілісно дослідити проблематику та визначити основні (соціо)лінгвістичні змінні. Визначено ключові моделі мовної варіативності когнітивних дієслів, зокрема тенденцію до розвитку абстрактних значень, відхилення від граматичної норми, а також ключові екстралінгвістичні чинники варіативності.

Ключові слова: мовна варіативність, мовна зміна, мовний варіант, мовна змінна, корпус.