Open Journal of Modern Linguistics, 2014, 4, 260-274
Published Online June 2014 in SciRes. http://www.scirp.org/journal/ojml
http://dx.doi.org/10.4236/ojml.2014.42021

Scientific Research

A Corpus Study on "Begin"/"Start" in Academic Writing: A VARBRUL Approach

Jia-Cing Ruan

Institute of Linguistics of National Chung Cheng University, Chiayi, Taiwan
Email: ac27037@gmail.com

Received 21 February 2014; revised 25 March 2014; accepted 3 April 2014

Copyright © 2014 by author and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

The present study aims to investigate the factors affecting the choice of aspectual verbs "begin" or "start", the factors affecting the choice of their complements, and the possible relation between the two choices in English Academic Writing. With the VARBRUL approach, a model has been found to account for these questions. The choice of "begin" or "start" is affected by the preceding syntactic tags, the choice of the complement is thus controlled by the choice of the aspectual verb, and both the two choices are affected by a social factor. In addition, the model of the VARBRUL analysis can be further applied to Natural Language Processing and English Academic Writing Teaching or Assistance, so a VARBRUL analysis is thus very economic. Other findings show the comparison of the model prediction among Logistic Regression, Decision-Tree-based Logistic Regression and Neural Network with linguistic data.

Keywords

Aspectual Verbs, VARBRUL, Academic Writing, Word Choice, Corpus

1. Introduction

The near-synonym pair "begin" and "start" (focusing on the function of verb) has been studying for at least twenty years. Mair (2003) has provided the historical descriptions of "begin" and "start" diachronically and synchronically, and has also provided good literature review of Biber, Conrad and Reppen (1998), and Biber et al. (1999) about the variation of "begin" and "start". Among the studies, "begin" and "start" have been studied in respect of the factors of the complementation (i.e. begin/start with gerund (such as "He starts/begins ***doing*** his homework.") or infinitive (such as "He starts/begins ***to do*** his homework."); henceforth, "ing-complement" or "to-complement in the present study") and of the registers and genres. Mair (2003) has studied different variants of English such as British and American English (More previous, Leitner (1994) had already studied three va-

How to cite this paper: Ruan, J.-C. (2014). A Corpus Study on "Begin"/"Start" in Academic Writing: A VARBRUL Approach. *Open Journal of Modern Linguistics, 4,* 260-274. http://dx.doi.org/10.4236/ojml.2014.42021

riants of English: British, American and Indian English). Leitner (1993) compared the usage of "begin" and "start" with different dictionaries. With Google Ngram Viewer (Lin et al., 2012), "begin" and "start" also show the interchangeable tendency with time progressing, as in **Figure 1** and **Figure 2**.

Figure 1 is the occurrences of "begin" and "start" of the past tense and perfect participate during 1800 to 2008, while **Figure 2** is the ones of "begin" or "start" of the base and third-person-singular forms; both queries are case sensitive. In **Figure 1** or **Figure 2**[1], the y-axis represents the proportion of the total corpus. Both **Figure 1** and **Figure 2**[2], especially **Figure 2**, show the much more increasing use of "start" in academic proses through comparing the difference between "begin" and "start" during 1800 and between the two during 2008. Even in the **Figure 1**, "begin" and "start" intends to have similar proportions, which implied the interchangeability. Moreover, as the results of WORDSKETCH (Kilgarriff & Tugwell, 2001) in **Table A1** (in Appendix), the inter changeability can thus be seen.

Previous studies focused more on the complementation of the aspectual verbs "begin" or "start" and their distributions in syntactic structures or different genre prose from corpora. The present study is trying to explain how "begin" or "start" is chosen and how their complementation relates to the "begin" or "start" choice. Once the model is accomplished, it is very contributive to extend to other research about aspectual verb choice as in Natural Language Processing, Machine Translation and so on.

The structure of this paper is as follows. Section Literature Review reviews studies about "begin" or "start". Section Methodology addresses the data source in the PERC corpus (Professional English Research Consortium, n.d.) and the statistical procedures (such as Rbrul (Johnson, 2009), CART Decision Tree (Breiman et al., 1984),

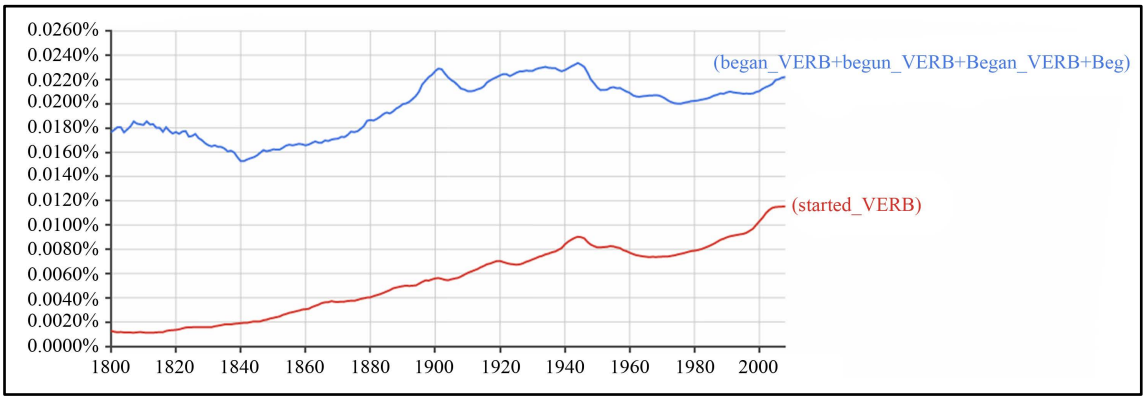

Figure 1. The diachronic occurrences of (began_VERB + begun_VERB + Began_VERB + Begun_VERB), (started_VERB).

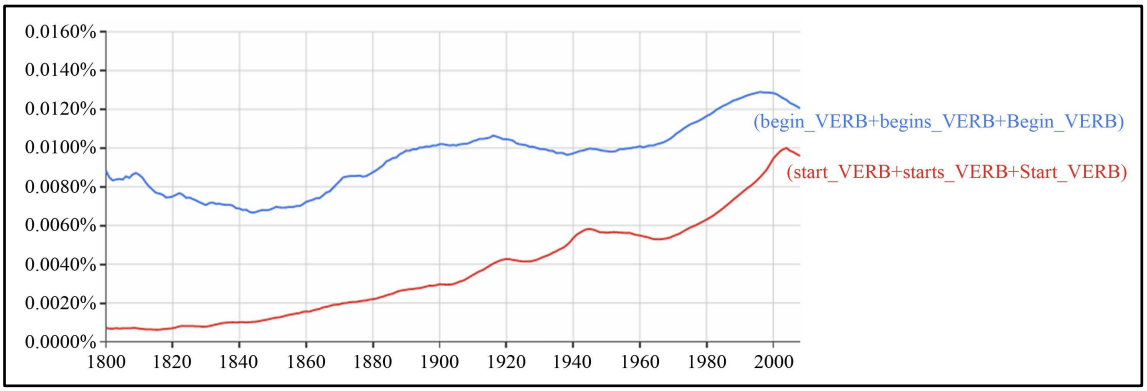

Figure 2. The diachronic occurrences of (begin_VERB + begins_VERB + Begin_VERB), (start_VERB + starts_VERB + Start_VERB).

[1]Accessed from Google Ngram Viewer on 2014/02/6: https://books.google.com/ngrams.

[2]In response to an anonymous reviewer, the Google Ngram Viewer has only one corpus for search on the website. Moreover, the **Figure 1** and **Figure 2** are used to show the possibility of interchangeability, so the detailed explanation is not the focus in the present study.

and Logistic Regression). Section Results offers the inferential statistical results. Section Discussion is the discussion of the results with the literature mentioned. Section Application addresses the application of VARBRUL (i.e. Variable Rule) analyses (a specialized Logistic Regression) to Natural Language Processing and English Academic Writing and assistance. Section Other Findings describes some findings of the relationship between the data and the statistical model in the present study. The last section is the conclusion, limitation and future development.

2. Literature Review

Freed (1979) has researched out that "begin" and "start" are semantically distinct. "start" expresses a stative point, while "begin" describes the initial duration. However, Dixon (2005) states that "begin" and "start" are interchangeable with less meaning lost.

Leitner (1993) has found that "begin" and "start" have different syntactic distributions with LOB corpus (Johansson, 1978). In the verb forms of tense and number, "begin" has the largest proportion of past-tense form, while "start" of verb-base and past-tense ones. In the eight syntactic patterns of the complement choice, "begin" has much more proportions of to-complement (261/489, 53.4%) than the ing-complement (24/489, 4.9%), while "start" has more of ing-complement (53/342, 15.5%) than the to-complement (37/342, 10.8%). In the use of valency, both "begin" and "start" are more frequently used as an intransitive (109/489, 22.3% and 116/342, 33.9%, respectively) than as a transitive (45/489, 9.2% and 104/342, 30.4%, respectively). As can be seen, "start" has small-scale difference (around 3%) between the use of intransitive and transitive; however, "begin" is much more used as an intransitive (around 13.1%). Therefore, "begin" and "start" behavior differently whatever in the complement choice or the use of valency.

Leitner (1994) has studied "begin" and "start" in different English varieties of British English (Brown corpus), American English (LOB corpus) and Indian English (Kolhapur corpus). His findings reveal distinct behaviors of "begin" and "start". "Begin" is used more frequently in written data, while "start" is more in spoken one. Leitner (1994) categorizes complements as: No complement, to-complement, ing-complement and NP-complement. The frequency rank of the four complements of "begin" is to-complement > No complement > NP complement > ing-complement, while the different rank of "start" is No complement > NP complement > ing-complement > to-complement. Among different varieties of English, the ranks of the four complements are different. Especially, "start" in American English has slightly more frequency with the to-complement, while "begin" has more with the ing-complement in British and India English. Leitner (1994) has also found that text categories (i.e. genres) influences the behaviors of "begin" or "start".

Biber, Conrad and Reppen (1998) have investigated "begin" and "start" in terms of valency (i.e. intransitive and transitive) and genres (i.e. fiction and academic). Both "begin" and "start" as intransitive have more proportions in academic genre than fiction. In their study, the valency of transitive has also been divided into three patterns: noun phrase, to-clause, and ing-clause. In the noun phrase and ing-clause categories, "begin" is preferred in academic, while "start" is in fiction. In the to-clause category, both "begin" and "start" are preferred in fiction. Corpus-based investigations are helpful to distinguish the near-synonyms, which even native speakers may fail to predict the difference[3].

Biber et al. (1999) has researched the ing-clause of "begin" and "start" in different registers: conversation, fiction, news and academic prose. The frequencies (per million words) of "start" in controlling the ing-clause in all registers are higher than "begin".

Mair (2003) has studied the complement choices of verbs "begin" and "start", comparing the differences between British English and American English, with several corpora. In both two varieties of English, to-complement is much more favored by "begin". On the other hand, Mair (2003) also has invited 31 British and 43 Americans to participate an informal experiment. The result shows that Americans somewhat more prefer the ing-complement after "begin" or "start", although both the British and Americans invited prefer the ing-complement.

Reiter and Sripada (2004) study on the contextual influences on near-synonym choice with semantic differences and social factors. They reveal that the impact to the choice is much more by social factors than semantic differences.

[3]Undoubtedly, a language use is tyranny of the majority. The benefits of corpus-based investigation are to show the overview usage from the performances in the society, which provides objective disambiguation for synonyms.

Szmrecsanyi (2006) has studied the complements of 11 base verbs with spoken data in English corpora. "start" is relatively, as in Biber, Conrad and Reppen (1998), more with the ing-complement. Furthermore, the stative complement after the 11 verbs decreases the odds of ing-complement in American English. However, the phenomena occurs only to the verbs of "begin" or "start" in British English. Szmrecsanyi (2006: p. 178) concludes "Horror aequi is an important determinant of complementation choice", and "begin" or "start" follows the principle.

Egan (2008) finds out "begin" outnumbers "start" in BNC corpus whatever with themselves or the to-complement. Furthermore, "start" is frequently used in conversations and is thus more in spoken data.

Mair (2009) has studied "begin" and "start" with corpora of different varieties of English. Within corpora: FLOB, ACE, WWC, and Frown, "begin" has more frequency followed by to-complement than ing-complement, while "start" is contrary. In addition, Mair (2009) has searched in ICE corpus with written and spoken English of different varieties of English: Great Britain, New Zealand, and Australia English. In each of the three varieties, "begin" has more frequency of to-complement than ing-complement, despite written or spoken data. Unusually to other results of previous literature, in both Great Britain and New Zealand English, the pattern of "begin" and "start" is identical that to-complement outnumbers ing-complement.

Gawlik (2012) has investigated how "begin" and "start" behavior with complement choices in spoken academic American English (MICASE corpus). In the academic lectures that MICASE recorded, "start" outnumbers "begin". "begin" is frequently used with the to-complement, while "start" is with the ing-complement.

However, less of them utilized inferential statistical procedures (especially the regression-like approach) to provide advanced analyses to reveal the reasons of choosing "begin" or "start" (i.e. the word choice) and its complements of "to-complement" or "ing-complement" (i.e. the complement choice). Most of the corpus-based studies provide only the descriptive statistics. The importance to apply an inferential statistics can be referred to provide more systematic analyses, more specific information and more applications to multiple fields whatever in Linguistics, Natural Language Processing and English teaching. Thus, this present study aims to carry out a variation study on the verb of "begin" and "start" with statistical procedures (especially the VARBRUL analysis (Labov, 1969; Rousseau & Sankoff, 1978)) to investigate the relationship between "begin/start" choice and its complement choice.

3. Methodology

The data used in the present study are from the PERC Corpus (Professional English Research Consortium). It is compiled with 17-million words from journal paper texts in English in 22 scientific fields (as in **Table 1**), mainly designed for Professional English research[4].

It provides the mark-up tagging of sentence boundaries, parts of speech and lemma. With these meta-marks, the corpus provides good information for statistical analyses.

For research of "begin" and "start" in the current study, the selection constraint of subcorpora was set to be the subcorpora with the size over 1,000,000 words for reducing the probability of effect of subcorpus with small sample size. **Table 2** is the summary table of selected corpora and its size.

Several statistical procedures were applied to analyze these data for a variation research: VARBRUL analysis (Variable Rule analysis), and Decision Tree (especially, CART (Classification and Regression Tree), (Breiman et al., 1984)).

In order to conduct the variation study, the usage environments of "begin" and "start" have been constrained to be identical, which can be presented as the diagram (**Figure 3**) below.

Section A: The usage environments never occur with "start".

Section B: The usage environments never occur with "begin".

Section C: The shared usage environments with both "start" and "begin".

Whatever in functions or usage environments of "begin" and "start", they are very easy to be disambiguated by the non-occurrences of each other, such as in A or B. However in C (i.e. the free variation), it is very difficult to disambiguate the two words. The necessity (or contributions) to disambiguate the two words in C can be referred to the issue of near-synonym word choice, which can be applied in Natural Language Processing to produce human-like sentences or to help modify a sentence to be more human-like in some texts of different styles such as formal writing in journal papers (emphasized in the present study) or others.

[4]http://scn.jkn21.com/~percinfo/.

Table 1. The summary of categories in the PERC corpus.

Category Name	File Count	Word Count	Percentage (%)
Construction & Building Technology	13	44,068	0.27
Environmental Sciences	10	47,991	0.3
Forestry	20	58,760	0.36
Telecommunications	23	114,506	0.71
Metallurgy & Metallurgical Engineering	29	114,728	0.71
Nuclear Science & Technology	46	159,271	0.98
Fisheries	69	215,920	1.33
Oceanography	64	237,466	1.46
General Science	68	289,761	1.79
Mathematics	67	310,834	1.92
Food Science	93	392,453	2.42
Civil Engineering	110	403,628	2.49
Materials Science	114	454,046	2.8
Electrical & Electronic Engineering	172	632,179	3.9
Agriculture	193	663,695	4.09
Earth Science	318	1,163,157	7.17
Physics	325	1,229,505	7.58
Biology	399	1,410,344	8.7
Computer Science	403	1,469,298	9.06
Chemistry	520	1,544,478	9.52
Engineering	518	1,929,699	11.9
Medicine	1125	3,330,238	20.54

Table 2. The summary of selected corpora in the present study.

Category Name	File Count	Word Count	Percentage (%)
Earth Science	318	1,163,157	7.17
Physics	325	1,229,505	7.58
Biology	399	1,410,344	8.7
Computer Science	403	1,469,298	9.06
Chemistry	520	1,544,478	9.52
Engineering	518	1,929,699	11.9
Medicine	1125	3,330,238	20.54

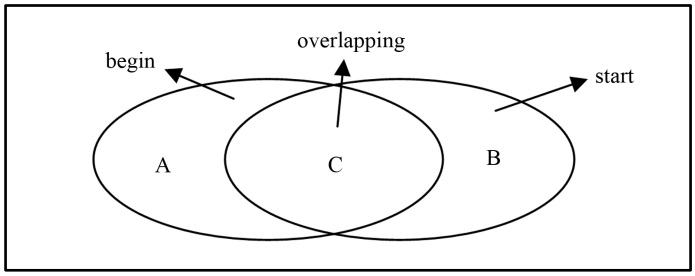

Figure 3. The diagram of usage environments of "begin" and "start".

In several references of word choice such as (Inkpen, 2009), their methods or algorithms are almost black-box, i.e. the attributes are unable to be examined in the model specifically. Additionally, some of the attributes used in the model or algorithm may be noise affecting the precision of classification. To avoid the defects mentioned and to provide a white-box model as a solution, the approach "VARBRUL analysis" can be applied to disambiguate near-synonyms with very good advantages in model tuning for predictions. Furthermore, it can also contribute to Natural Language Processing, linguistics fields (e.g. Semantics, Sociolinguistics) or English teaching (e.g. English writing) simultaneously.

A sociolinguistic variation study is usually conducted with the VARBRUL analysis through the software of GOLDVARB (Rand & Sankoff, 1990; Robinson, Lawrence, & Tagliamonte, 2001; Sankoff, Tagliamonte, & Smith, 2005). However in recent years, it seems to be a tendency from using GOLDVARB to GLM for some reasons as discussed in (Saito, 1999; Young & Yandell, 1999). Due to the different algorithms used in GOLDVARB and GLM, GOLDVARB can only accept frequency data in both dependent variable and independent variables, which is thus criticized. The present study utilized Rbrul to conduct the VARBRUL analysis for its convenience.

Rbrul (Johnson, 2009) package in R statistics (R Core Team, 2013) provides an eclectic way for analyses linking GOLDVARB and GLM. Rbrul provides a correspondent transformation between log-odds produced by GLM and the VARBRUL weights (i.e. a probability of a factor) for conducting a VARBRUL analysis. Due to the use of "glm" package in R, the Rbrul transforms the log-odds to VARBRUL weights regardless of the assumption of no interaction between independent variables.

The reason of using Rbrul rather than GOLDVARB for a variation study is explained. In the present study, the factors in independent variables are very numerous. GOLDVARB requires users to recode the labels of factors to one-digit scheme, and disallows 100% or 0% occurrence of each factor in independent variables. When facing 100% or 0% occurrence, GOLDVARB cannot run for analyses. Without the inconveniences, Rbrul provides more flexible acceptance of no recode labels to one-digit scheme and automatically handle the 100% or 0% occurrence.

On the other hand, Decision Tree (CART), a machine learning approach, was applied to automatically reveal the interactions between independent variables through the "rpart" package (Therneau, Atkinson, & Ripley, 2013) in R, since Rbrul or Logistic Regression of GLM cannot automatically produce the results. It is often used to classify samples with multi-attributes.

Research conducted with GOLDVARB software is restricted with frequency data for both dependent variable and independent variables. However, Rbrul and CART can both be used with non-frequency data in independent variables.

The dependent variable is the use of free variation switched by "begin" and "start". As mentioned, their usage environments are restricted as the same ones. As for the independent variables, there are four independent variables. Three of them are in the usage environments (the preceding POS tag, the inflection of "begin" or "start", and the following POS tag), and the other is the different domains selected.

In order to reduce the data sparseness (i.e. the factor levels in independent variables) during the analyses, the tags used in PERC corpus were reduced and reclassified to a set of more general categories as provided in PERC[5], where, for example, NN0, NN1, NN2 and NP0 were recategorized to Noun. In the following section, the results are presented.

[5]https:/scn.jkn21.com/~perc04/bncposA2.html.

4. Results

4.1. Results of Main-Effects

In order to provide a clear presentation of results, the analyses are reported in two sections for main-effect analyses (assuming no interaction) and for interaction-effect analyses. The first part deals with the Rbrul analysis with main effects only, and the second part presents not only the interaction effects by Decision Tree (CART), but also the significance test and the best model. The reason of significance test for Decision Tree (CART) is to ensure the event probabilities in the model as discussed in (Eddington, 2010). Namely, the interaction variables found by Decision Tree (CART) was used in a Logistic Regression model for significance test.

In the One-level (i.e. full enter method) analysis of main effects (**Table 3** to **Table 6**), all the independent variables are statistically significant with Input Probability 0.471 by "begin". The interpretation of the VARBRUL weights concerns about the value equivalent to 0.5 or over 0.5 (i.e. favor), and below 0.5 (i.e. disfavor). In the

Table 3. The VARBRUL analysis of the independent variable BEFORE[6].

BEFORE, $p = 0.03$				
Factor level	logodds	tokens	begin/begin + start	centered factor weight
Infinitive marker to	0.754	139	0.511	0.68
Modal auxiliary verb	0.752	79	0.557	0.68
Numeral	0.304	13	0.615	0.576
Verb	0.278	219	0.507	0.569
Pronoun	0.214	237	0.570	0.553
Adverb	0.181	126	0.603	0.545
Conjunction	0.040	97	0.588	0.51
Punctuation	−0.017	180	0.411	0.496
Noun	−0.168	1042	0.531	0.458
Wh-determiner-pronoun	−0.341	22	0.455	0.416
Others	−0.348	24	0.542	0.414
Sentence initial	−0.701	65	0.215	0.332
Preposition	−0.949	73	0.192	0.279

Table 4. The VARBRUL analysis of the independent variable INFLECTION.

INFLECTION, $p < 0.01$				
Factor level	logodds	tokens	begin/begin + start	centered factor weight
VVD[7]	0.497	593	0.626	0.622
VVZ	0.337	488	0.527	0.584
VVB	0.332	240	0.562	0.582
VVG	−0.173	551	0.356	0.457
VVI	−0.415	233	0.519	0.398
VVN	−0.579	211	0.474	0.359

[6]In the present study, an independent variable refers to a factor.

[7]VVD is for "past tense", VVZ for "-s form", VVB for "base form", VVG for "-ing form", VVI for "infinitive form" and VVN for "past participle".

Table 5. The VARBRUL analysis of the independent variable AFTER.

AFTER, $p < 0.01$				
Factor level	logodds	tokens	begin/begin + start	centered factor weight
Infinitive marker to	0.678	496	0.696	0.663
Pronoun	0.650	17	0.706	0.657
Punctuation	0.292	148	0.608	0.573
Conjunction	0.117	57	0.579	0.529
Verb	−0.078	89	0.584	0.481
Numeral	−0.261	44	0.500	0.435
Preposition	−0.284	1170	0.426	0.429
Noun	−0.323	48	0.438	0.42
Adverb	−0.387	86	0.465	0.404
Adjective	−0.404	161	0.416	0.4

Table 6. The VARBRUL analysis of the independent variable CATEGORY.

CATEGORY, $p < 0.01$				
Factor level	logodds	tokens	begin/begin + start	centered factor weight
Medicine	0.498	579	0.596	0.622
Biology	0.491	276	0.620	0.62
Earth	0.305	235	0.570	0.576
Computer	0.056	472	0.500	0.514
Engineering	−0.307	324	0.414	0.424
Chemistry	−0.403	242	0.393	0.401
Physics	−0.640	188	0.346	0.345

BEFORE environment, "begin" favors categories preceding: "Infinitive marker to", "Modal auxiliary verb", "Numeral", "Verb", "Pronoun", "Adverb", and "Conjunction". The others are disfavored by "begin", but favored by "start". In the INFLECTIONAL, "begin" favors forms of VVD, VVZ, and VVB, but disfavors VVG, VVI and VVN. In the AFTER environment, "begin" favors "Infinitive marker to", "Pronoun", "Punctuation" and "Conjunction", but disfavors the others. In CATEGORY, "begin" is used favoritely in journal domains of "Medicine", "Biology", "Earth" and "Computer". Factor levels which are disfavored by "begin" are favored by "start".

A step up/down analysis was performed to reach the best model of the analysis. All the independent variables were remained as significant. In other words, the One-level analysis is the best model.

4.2. Results of Interaction-Effects

Detecting interactions under no hypotheses in a Logistic Regression model could be very vexing due to no automation. Researchers may usually grab all possible interaction sets as a model for Logistic Regression analyses, where significant interactions are reported. The shortcomings and risks would be very time-consuming (if the factors or levels are many and with too much sparse, and would be no convergence in the algorithm). Below is the saturated model (i.e. all possible interaction sets) with 15 hours waiting[8].

[8]CPU: AMD Athlon 64 × 24,000+, RAM: 3 GB, Microsoft Windows XP 32 bit. It will be faster in Ubuntu 64 bit (only about 5 hours) under the same environment.

	Df	Deviance	Resid. Df	Resid. Dev	Rao	Pr(>Chi)	
NULL			2315	3210			
before	12	76	2303	3134	7.3000e+01	9.878e-11	***
inflection	5	54	2298	3080	5.3000e+01	2.691e-10	***
after	9	70	2289	3010	6.9000e+01	2.106e-11	***
category	6	73	2283	2937	7.2000e+01	1.336e-13	***
before:inflection	22	27	2261	2910	2.7000e+01	0.218631	
before:after	44	45	2217	2865	4.3000e+01	0.526618	
inflection:after	18	17	2199	2848	1.6000e+01	0.606310	
before:category	68	130	2131	2718	1.1000e+02	0.001003	**
inflection:category	29	0	2102	66969	5.7000e+01	0.001454	**
after:category	53	64398	2049	2571	1.7304e+18	< 2.2e-16	***
before:inflection:after	11	9	2038	2562	9.0000e+00	0.622908	
before:inflection:category	67	0	1971	59544	7.0000e+01	0.387772	
before:after:category	111	0	1860	78287	1.7980e+18	< 2.2e-16	***
inflection:after:category	52	22275	1808	56012	3.0703e+18	< 2.2e-16	***
before:inflection:after:category	6	0	1802	81603	1.7270e+18	< 2.2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In the saturated model, several interactions are reported as significant: "before:category", "inflection:category", "after:category", "before:after:category", "inflection:after:category", and "before:inflection:after:category". The result reveals the four statistically significant independent variables interact with others. Especially, the social variable "CATEGORY" much more interacts with linguistic variables "BEFORE, AFTER, or INFLECTION", as emphasized in Sigley (2003). Therefore, selecting "begin" or "start" is not only caused by environments (i.e. BEFORE or AFTER), but the interactions among environments, inflections of "begin" or "start" (i.e. INFLECTION) and different domains of journal papers (i.e. CATEGORY).

Regard of Sigley (2003), the analyses of the interaction effects are very beneficial in variation studies. However, detecting and probing the interactions between independent variables manually are very time-consuming, as in the example of the Logistic Regression analysis above. Decision Tree (especially CART, a machine learning method) was applied to automatically reveal the interactions efficiently. The results are as follows.

In **Figure 4**, rules can be elicited as follows (partial):

1) If AFTER is "Cnjn", "Inmt", "Prnn", "Pnct", or "Verb", then select "begin".

2) If AFTER is "Adjc", "Advr", "Noun", "Nmrl", or "Prps" then select "start".

3) If AFTER is "Adjc", "Advr", "Noun", "Nmrl", or "Prps" and INFLECTION is "VVB", "VVD", "VVI", or "VVZ", then select "begin".

4) If AFTER is "Adjc", "Advr", "Noun", "Nmrl", or "Prps" and INFLECTION is "VVG" or "VVZ", then select "start".

For example, if AFTER is "Noun", INFLECTION is "VVB" and CATEGORY is "Earth", then the selection of the dependent variable is "begin". In VARBRUL words, "begin" favors in the form of "VVB", followed by "Noun", and in the "Erth" domain of journal papers. As known, Decision Tree selects the most valued model according to the data; therefore, BEFORE is not a very important variable in the Decision Tree analysis. The interaction sets found by Decision Tree is "after:inflection", and "after:inflection:category". As in Eddington (2010), Decision Tree is not as very precise as in a Logistic Regression model. However, the interaction sets can still be thrown into a Logistic Regression model for a significance test. Different from Eddington (2010), the present study put the interaction sets with main effects to form a Logistic Regression model. Below is the result of significance test. Both "after:inflection" and "after:inflection:category" are statistically significant.

	Df	Deviance	Resid. Df	Resid. Dev	Rao	Pr(>Chi)	
NULL			2315	3210			
before	12	76.220	2303	3134	72.723	9.878e-11	***
inflection	5	53.738	2298	3080	53.472	2.691e-10	***
after	9	70.243	2289	3010	69.281	2.106e-11	***
category	6	72.953	2283	2937	72.366	1.336e-13	***
inflection:after	33	57.053	2250	2880	54.946	0.009607	**
inflection:after:category	206	0.000	2044	64014	263.794	0.004028	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The main-effect analyses can be discussed in respect to internal factors (i.e. linguistic factors: BEFORE, INFLECTION, AFTER) and external factors (i.e. social factors: CATEGORY) in the present study.

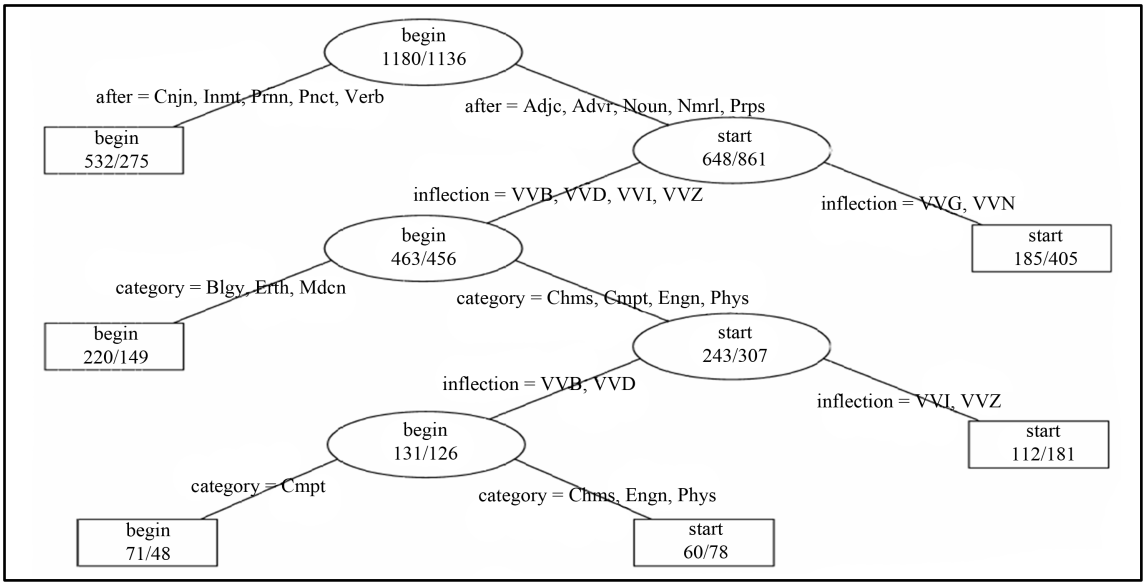

Figure 4. The classification tree for begin/start.

5. Discussion

5.1. Internal Factors

With internal factors, the main-effect analyses show that "begin" relatively prefers the to-complement while "start" prefers the ing-complement, which is consonant with the previous studies (Mair, 2003; Biber, Corad, & Reppen, 1998; Biber et al., 1999; Leitner, 1993; Leitner, 1994; Mair, 2009; Gawlik, 2012; Szmrecsanyi, 2006). "start", relatively to "begin", favors NP-complement, as in Leitner (1994). Both "begin" and "start" favor past tense inflection, similar to (Leitner, 1993).

The reason of the distinct preferences might be ***indirectly relevant*** to the Horror aequi principle (Szmrecsanyi, 2006). Analogized with Horror aequi principle, the final syllable "nucleus (high, front) + coda (nasal)" of "begin" might disprefer the similar final syllable of any ing-complements ("nucleus (high, front) + coda (nasal)"). Accordingly, the coda of "start" keeps dissimilar to the onset of the to-complement[9]. Thus, "begin" or "start" determines the to-complement or the ing-complement—despite that BEFORE and AFTER are two significant factors. However, the choice of to-complement or ing-complement is not that absolute for "begin" or "start".

The previous studies focus more on the choice of the complement; however, the present study more emphasizes on the choice of "begin" or "start", which seems not to be paid attention. All the usage environments (BEFORE, INFLECTION, and AFTER) are statistically significant factors influencing the choice of "begin" and "start". According to the strength rank of influencing the choice of "begin" and "start" (BEFORE > INFLECTION, BEFORE > AFTER, INFLECTION = AFTER), BEFORE is the most important factor. It seems that "begin" relatively neighbors more functional categories (i.e. infinitive marker, model auxiliary, or adverbs). Additionally, less nouns neighbor before "begin"; instead, pronouns do. From the observation above, it seems to be able to attribute categories with semantic function (i.e. possibility, or modification) to determine the choice between "begin" and "start".

5.2. External Factor

In the main-effect analyses, CATEGORY is the only one external factor with statistical significance in the present study. It consists of different academic fields accounted as one kind of genre. As in Biber, Conrad and Reppen (1998), "begin" and "start" distribute differently in genres (i.e. fiction, and academic). As in Biber et al. (1999), they distinctly distribute in conversation, fiction, news and academic prose. The choice of "begin" or

[9]The Horror aequi principle not only shows the avoidance of two adjacent structures which are similar, but also two segments such as the phonological Horror aequi effect in Gries and Hilpert (2010).

"start" is affected by the different academic fields.

The main-effect analyses assume there is no interaction between any main effects. However, Sigley (2003) depicts the importance of interaction analyses that can reveal or discover more information and results. The following section discusses the interaction-effect analyses.

Referred by Sigley (2003), the interaction between internal factors and between internal and external factors are two directions for analyses and interpretations. Both the Decision-Tree-based Logistic Regression model and the saturated Logistic Regression one indicate of "inflection:after:category" as a significant interaction effect. However, the saturated model shows insignificant "inflection:after", and significant "inflection: category" and "after:category". From the observations above, CATEGORY seems to attribute the most important role to the significant "inflection:after:category". In addition, all the significant interaction effects in the saturated model are interacted with CATEGORY, which also indicates that all the internal-external interactions are significant. Moreover, from the saturated model, all the internal-internal interactions are insignificant. Thus, it is the preferred usages that influences "begin" and "start" in the academic papers of different scientific domains, which is supported that the social factors have more influences than linguistic factors (Reiter & Sripada, 2004).

This section discussed about the main-effect and interaction-effect analyses. "begin" and "start" distributively share the usage environments with the impacts of the social factor. The interactions between internal and external factors are very evident. The summarized model of determining "begin" or "start" with the to-complement and the ing-complement is illustrated as **Figure 5** (arrows stand for determination):

6. Application

In this section, VARBRUL analyses are shown to have feasible applications in Natural Language Processing and English Writing Assistance. With VARBRUL weights and factor ranks, VARBRUL results may be utilized as an extension accompanying Decision Tree, featuring the hierarchy determination and the significance test.

In the model of **Figure 5**, the most influential factor affecting the choice of "begin" or "start" is BEFORE, and the choice of the to-complement or the ing-complement is affected by the choice of "begin" or "start" (i.e. Horror aequi principle). However, these factors are all impacted (or interacted) by CATEGORY. The VARBRUL weights in this model may be used similarly to the back-off in N-gram to elicit rules. Moreover, when the tokens of multiple-interaction is too less to explain, it can back to the level of multiple-minus one-interaction. Tackling the problem of decision of priority, the factor rank can contribute to the model. Following is the partial[10] example of the interaction effect "after:category" in the model:

Dependent-before + inflection + after + category + after:category.

In **Table 7**, these factor levels are preferred by "begin". From the first level, for example, a rule can be thus elicited as: "begin" prefers to be after "punctuation" in the academic writing in "Chemistry" domain.

As the back-off, one factor level has no token such in "pronoun:physics" of "start", which can back to "pronoun" in AFTER or "physics" in CATEGORY and the importance can be determined by the model hierarchy (i.e. the comparison of the strength of AFTER and the one of CATEGORY).

In comparison of Decision Tree, VARBRUL results provide information that is more specific and is more flexible. The VARBRUL weights and factor ranks are absent in Decision Tree. Moreover, Decision Tree is relatively rough in its analyses (Inkpen, 2007). However VARBRUL analyses are inappropriately conducted in small sample size, with lower efficiency and without automatic interaction detection. Therefore, regardless of efficiency and sample size, VARBRUL analyses may still provide more precise relative probabilities to each factor level and the hierarchy among factors.

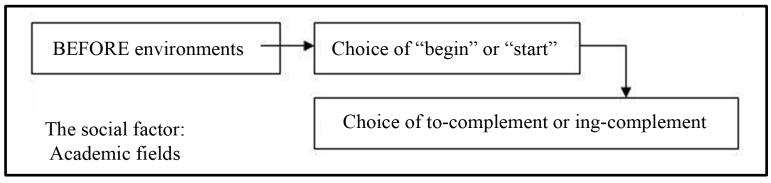

Figure 5. The progressive model for the choice of "begin" or "start" and complements.

[10]The whole report is too long to display.

Table 7. The interaction effect of "after:category" preferred by "begin".

Factor level (after:category)	logodds	tokens	begin/begin + start	centered factor weight
Punctuation:Chemistry	4.259	10	0.3	0.986
Conjunction:Engineering	4.106	8	0.25	0.984
Preposition:Medicine	3.642	257	0.494	0.974
Adjective:Engineering	3.603	27	0.333	0.973
Infinitive marker to:Engineering	3.348	65	0.585	0.966
Infinitive marker to:Medicine	2.673	125	0.832	0.935
Adjective:Medicine	2.65	48	0.521	0.934
Conjunction:Chemistry	2.256	3	0.667	0.905
Noun:Engineering	2.048	4	0	0.886
Preposition:Biology	2.029	100	0.57	0.884

VARBRUL analyses may be further applied to English writing teaching or English writing assistance software. Generalizing to English teaching with the model in **Figure 5**, English writing instructors may analogize to the word choice of other synonymy sets to teach the interrelations between English verbs and usage environments. More extensively, English writing instructors may remind different preferences in different genres (such like different scientific fields in the present study) with VARBRUL analyses. The realization of VARBRUL analyses to a program or a plugin as an English writing assistance in text editor software is thus feasible. The writing aid provides academic paper authors, for example, information about preferred word choices in genres.

7. Other Findings

During conducting VARBRUL analyses, the present study observed that the relation between the data and the model is mutually influential. Studies in Natural Language Processing (especially the word choice) concentrate more on developing models from statistics or other techniques. Their goals are to develop a good model to fit language data as complete as possible. However, the labels of data (i.e. data attributes or features) are, on the other side, important as well. The appropriate attributes or features plus a well-fit model are the key to well explain language data. In the data of the present study, it was found that the precision of the model prediction could be enhanced through interactivizing data attributes or features (i.e. a saturated model) without changing the model. Presented in the following, Logistic Regression model, Decision-Tree-based Logistic Regression model and Neural Network are compared on the precision of model prediction, with the tool SPSS.

As can be seen in **Table 8**, models with interaction effects provide more precise prediction than the models only with main effects. Logistic Regression achieves a little more precise predictions than Neural Network whatever in the main effect model or the saturated models. However, Logistic Regression spent too much time and thus is very inefficient. As for the Decision-Tree-based Logistic Regression, it provides only fair precision compared with the saturated models.

From the findings, the precision of model prediction seems never unidirectional. It is the mutual best-match between the data and model that compromises the goodness-of-fit.

8. Conclusion

The present study has investigated "begin" or "start" with the VARBRUL analyses. It contributes a model accounting for the choice of "begin" or "start" and the choice of "ing-complement" or "to-complement". Furthermore, it contributes elicited rules through the VARBRUL approach to be company with Decision Tree in Natural Language Processing. Moreover, it contributes a feasible implementation aiding English academic writing or writing pedagogy. With the VARBRUL approach, a study can simultaneously contribute results to multi-fields; thus, it is a very economic way to spend less time and harvest much more fruitful products.

Table 8. The comparison of models for predict "begin" or "start".

Method	Model	Sample	Precision	Time Spent
Logistic Regression	main effects[a]	2316	63.6	Less than 1 min
Neural Network	main effects[a]	2316	66.31	Less than 1 min
Decision-Tree-Based Logistic Regression	main effects + interaction effects[b]	2316	69	Less than 1 min
satuated Logistic Regression	main effects + interaction effects[d]	2316	73.1	About 15 hours
satuated Logistic Regression	main effects + interaction effects[c] (Significant Only)	2316	72.5	About 15 hours
Neural Network	main effects + interaction effects[d]	2316	72.3	5 mins

Note: [a]before + inflection + after + category; [b]before + inflection + after + category + inflection:after + inflection:after:category; [c]before + inflection + after + category + before:category + after:category + category:inflection + after:before:category + after:category:inflection + after:before:category:inflection; [d]before + inflection + after + category + before:inflection + before:after + inflection:after + before:category + after:category + before:inflection:after + before:inflection:category + category:inflection + after:before:category + after:category:inflection + after:before:category:inflection.

Several limitations are expected to progress. The data used in the present study are only written texts. Different varieties of English are not considered. Less external factors (i.e. social factors) are involved. Moreover, the model in the present study is much more intended to a syntactic analysis. Future studies will be expected to utilize semantic tags in possible corpora.

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511804489

Biber, D., Johansson, S., Leech, G., Conrad, S., & Fineqan, E. (1999). *Longman Grammar of Spoken and Written English.* Harlow, Essex: Pearson Education.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees.* Belmont, CA: Chapman and Hall.

Dixon, R. M. W. (2005). *A Semantic Approach to English Grammar* (2nd ed.). Oxford: Oxford University Press.

Eddington, D. (2010). A Comparison of Two Tools for Analyzing Linguistic Data: Logistic Regression and Decision Trees. *Italian Journal of Linguistics, 22,* 265-286.

Egan, T. (2008). *Non-Finite Complementation: A Usage-Based Study of Infinitive and -ing Clauses in English.* Amsterdam: Rodopi.

Freed, A. F. (1979). *The Semantics of English Aspectual Complementation.* Dordrecht: Springer. http://dx.doi.org/10.1007/978-94-009-9475-1

Gawlik, O. (2012). On the Complementation of Start, Begin and Continue in Spoken Academic American English. *Token: A Journal of English Linguistics, 1,* 159-170.

Gries, S. T., & Hilpert, M. (2010). Modeling Diachronic Change in the Third Person Singular: A Multifactorial, Verb- and Author-Specific Exploratory Approach. *English Language and Linguistics, 14,* 293-320. http://dx.doi.org/10.1017/S1360674310000092

Inkpen, D. (2007). A Statistical Model for Near-Synonym Choice. *ACM Transactions on Speech and Language Processing, 4,* 2.1-2.17. http://dx.doi.org/10.1145/1187415.1187417

Johansson, S. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.* Oslo: Department of English, University of Oslo.

Johnson, D. E. (2009). Getting off the Goldvarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass, 3,* 359-383. http://dx.doi.org/10.1111/j.1749-818X.2008.00108.x

Kilgarriff, A., & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of the ACL Workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation* (pp. 32-38). Toulouse: Association for Computational Linguistics.

Labov, W. (1969). Contraction, Deletion, and the Inherent Variability of the English Copula. *Language, 45,* 715-762. http://dx.doi.org/10.2307/412333

Leitner, G. (1994). Begin and Start in British, American and India English. *Hermes, Journal of Linguistics, 13,* 99-122.

Leitner, G. (1993). Where to Begin or Start? Aspectual Verbs in Dictionaries. In M. Hoey (Eds.), *Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair on His Sixtieth Birthday* (pp. 50-63). London: HarperCollins.

Lin, Y., Michel, J., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, (pp. 169-174). Jeju Island, Korea: Association for Computational Linguistics.

Mair, C. (2003). Gerundial Complements after Begin and Start: Grammatical and Sociolinguistic Factors, and How They Work against Each Other. In G. Rohdenburg &, B. Mondorf (Eds.), *Determinants of Grammatical Variation in English* (pp. 329-346). Berlin, New York: Mouton de Gruyter.

Mair, C. (2009). Infinitival and Gerundial Complements. In P. Peters, P. Collins, & A. Smith (Eds.), *Comparative Studies in Australian and New Zealand English: Grammar and Beyond* (pp. 263-276). Amsterdam: John Benjamins.

Professional English Research Consortium (n.d.). *PERC Corpus.* http://www.perc21.org/corpus_project/index.html

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. http://www.R-project.org/

Reiter, E., & Sripada, S. (2004). Contextual Influences on Near-Synonym Choice. *Proceedings of the International Natural Language Generation Conference*, Brockenhurst, 14-16 July 2004, 161-170. http://dx.doi.org/10.1007/978-3-540-27823-8_17

Rousseau, P., & Sankoff, D. (1978). Advances in Variable Rule Methodology. In D. Sankoff (Eds.), *Linguistic Variation: Models and Methods* (pp. 57-69). New York: Academic Press.

Saito Sigley, H. (1999). Dependence and Interaction in Frequency Data Analysis in SLA Research. *Studies in Second Language Acquisition, 21,* 453-475.

Sigley, R. (2003). The Importance of Interaction Effects. *Language Variation and Change, 15,* 227-253. http://dx.doi.org/10.1017/S0954394503152040

Szmrecsanyi, B. (2006). *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis.* Berlin, New York: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110197808

Therneau, T., Atkinson, B., & Ripley, B. (2013). *Rpart: Recursive Partitioning.* R Package Version 4.1-3. http://CRAN.R-project.org/package=rpart

Young, R., & Yandell, B. (1999). Top-Down versus Bottom up Analysis. *Studies in Second Language Acquisition, 21,* 477-488. http://dx.doi.org/10.1017/S0272263199003058

Appendix

Table A1. The WORDSKETCH analysis for "begin" and "start".

Function	Begin	Start	Function	Begin	Start	Function	Begin	Start
object	4913	7246	subject	18,301	10,589	modifier	4538	4854
engine	**0**	**113**	**recipe**	**0**	**84**	**afresh**	**0**	**38**
fire	**0**	**138**	**Sale**	**0**	**66**	**better**	**0**	**63**
Monday	**0**	**44**	**walk**	**0**	**28**	**somewhere**	**0**	**32**
favourite	**0**	**30**	**engine**	**0**	**30**	all	35	98
fight	**0**	**32**	**sale**	**22**	**79**	tomorrow	25	68
December	**0**	**33**	fire	21	66	again	233	491
business	**0**	**152**	train	23	36	today	36	71
season	24	68	trouble	44	59	actually	66	115
work	266	480	season	48	59	early	58	87
proceedings	57	59	universe	25	20	suddenly	45	62
campaign	71	67	work	227	188	then	253	291
process	130	98	rain	45	28	just	422	417
journey	41	29	war	119	68	immediately	107	84
career	180	111	story	72	35	now	295	224
search	36	23	process	92	47	soon	112	86
tour	47	27	career	56	24	even	195	125
negotiation	39	21	heart	72	28	already	311	178
discussion	52	26	**sun**	**43**	**0**	slowly	77	41
investigation	63	26	**trial**	**47**	**0**	**gradually**	**36**	**0**
descent	**31**	**0**	**talk**	**54**	**0**	**last**	**38**	**0**

Search Condition:
1) Frequency: over 20 counts.
2) Display: 20 items.
3) Part of Speech: Verb.

Interpretation:
1) Items in dark without square tend to be with "start", strength: 2 of 6.
2) Items in dark with square tend to be with "begin", strength: 2 of 6.
3) Others are no difference.