Empirical Regression of Entropy & Multivariate Mutual Information of Concgram-based Multi-Word Expressions on Learner Corpus

A Research Report Presented to the Department of Linguistics and Translation

College of Liberal Arts and Social Sciences

City University of Hong Kong

In Partial Fulfillment of the Requirements of the Degree of

MASTER OF ARTS

in

LANGUAGE STUDIES

(GENERAL LINGUISTICS)

by

LAU, Chan Wing Eddie

(Final Version)

Submitted on: 4 August 2017

Revised on: 15 September 2017

Table of Contents

Acknowledgement

Nurtured by the superb supervision of corpus and computational linguist Dr Alex Chengyu Fang (Associate Professor of the Department of Linguistic and Translation at City University of Hong Kong), this report's author must express his wholehearted and wordless gratitude to Dr Fang for inspiring the author of the present research with his immense wisdom, including the introduction of state-of-the-art papers. These recommendations profoundly help the present author settle down to productive work. He also allows ample room for the author to decide on some critical matters related to this research project while rectifying its direction if it goes wrong.

In addition, the author is indebted to the Run Run Shaw Library at this University as it facilitates this master's research by offering a package of the International Corpus of Learner English (ICLE) and by promptly acquiring phraseological search engine ConcGram©, designed and developed by Chris Greaves, an ex-Senior Research Fellow of The Hong Kong Polytechnic University. Furthermore, the Hong Kong Academic Library Link (HKALL) provides the present author with excellent joint local university borrowing services.

Equally important, the author must also be thankful for the excellent administration of and ongoing communication for the Master of Arts in Language Studies (MALS). Special thanks are due to Assistant Head of the Department of Linguistics and Translation & Program Leader of MALS Dr Richard Sheung, Subject Leader of MALS (Linguistics) Dr Wei Zhang, and their colleagues of the departmental general office for having arranged and approved the present author's initial proposal for this research project.

Spiritually, without the eternal, tacit yet tangible, love from parents, sisters and best friends, the author absolutely cannot launch and then round off the present master's research, a process in which a sense of loneliness sometimes creeps up on him although a sense of fulfilment also pops into his mind.

Finally, thanksgiving to God for arranging this miraculous, amazing and blessed endeavour pursued at the author's alma mater, CityU, and supervised by Dr Fang, not only eminent but also compassionate towards me.

Abstract

Quantitative linguistics on learner corpus research is the realm of the present research. To avoid vague research questions and non-testable hypotheses contrived without empirical grounds at the outset, an overarching bottom-up approach first identifies three linguistic phenomena—learner language, multi-word expressions (MWEs), and concgrams, i.e. phraseological constituency (e.g. "AB, A*B, A**B, etc.") and phraseological positional variation (e.g. "AB, BA, B*A, etc."). A concgram can be contiguous (i.e. n-gram), non-contiguous (i.e. skipgram), and hybrid, e.g., "after all", "both…and…", and "as much…as" respectively. Concgram-based MWEs are extracted from the International Corpus of Learner English (ICLE) by phraseological search engine ConcGram©.

But ConcGram© can generate only univariate significance tests and bivariate mutual information of words. The present research fills this much-needed gap—multivariate association among words in concgram-based MWEs—by adopting what we coin Markov multivariate mutual information (MMMI) devised by linguists Wei & Li (2013). Our novelty is to design a corpus-driven empirical multiple linear regression model which measures how dependent variable Shannon entropy, quantifying uncertainty, is correlated with or explained by independent variables MWE length and MMMI. Only then can research questions and hypotheses be laid down. Next, the regression model is built and tested using statistical computing language R.

The statistically significant finding is that independent variable MMMI tends to inversely relate to Shannon Entropy such that the larger the MMMI (i.e. information of association), the smaller the Shannon Entropy (i.e. uncertainty). And independent variable MWE Length tends to have a positive relation to Shannon Entropy such that the larger the MWE Length, the larger the uncertainty too. These two relations seem contradictory. But MMMI cannot be compared with MWE Length as the former's formulation is complicated whereas the latter's is simple—with values 2, 3, 4 and 5 only.

The two relations above are statistically significant in terms of the estimates and *p*-values of the two explanatory variables. What are often missed in reporting results are the checking for outliers and overmuch influential observations, and the verification of the assumptions of multiple linear regression. These checking and verification are particularly detailed in the present research. The overall goodness of fit R^2, which ranges from 0 to 1, is 0.6053, a quite high level; and, the F-statistic is 236.9, which is very high, and its *p*-value is very low so the null hypothesis—the explanatory variables have no effect—can be rejected. Thus, it is concluded that the model is significant. However, some of the assumptions of multiple linear regression are only marginally met or even marginally violated. This implies that the estimates might be unreliable. It is tentatively suggested that semiparametric regression, a remedial methodology, is to be adopted in future research.

Introduction

The present research begins with section Corpora and Empirical Approaches, which elucidates corpus-based plus corpus-driven approaches, the quantitative turn in learner corpus research (LCR), and an overarching bottom-up approach. Next, section Linguistic Phenomena with Literature Review appraises the literature on 3 phenomena: learner language, MWEs, and concgram-based MWEs—evidenced by the International Corpus of Learner English (ICLE). Then, section Research Methodology with Literature Review evaluates the literature on: multivariate mutual information (MMI) as an association measure, Shannon entropy as an uncertainty measure, multiple linear regression, spreadsheeting plus statistical computing with R, and a computational approach to concgramming. With these methodologies, research questions and hypotheses can then be set out. Afterward, the research enters section Empirical Analytics and Discussion as well as section Future Research. Finally, the research ends with section Conclusion and then a Bibliography.

Phraseological search engine ConcGram©, devised by Greaves (2009), extracts *concgram-based* MWEs that are contiguous (e.g. "as a matter of fact"), non-contiguous (e.g. "so…that") or hybrid (e.g. "not only…but also") from ICLE, which is overviewed by Granger, Dagneaux, Meunier & Paquot (Eds.) (2009). ICLE consists of 16 sub-corpora of texts produced by learners of English from 16 different countries of 16 L1s. Six of the sub-corpora contain *100% argumentative* essays, which are merged into an aggregate sub-corpus comprised of *1.12 million word-tokens* as primary data in the present study. As these essays are argumentative, factor *'genre'* is deliberately controlled.

To quantify the association among words within an MWE, Wei & Li (2013 pp. 508) trial a generalization of Shannon's bivariate mutual information (BMI) to multivariate mutual information (MMI), which integrates *pseudo-bigram*, a concept of Silva & Lopes (1999), with stochastic process *first-order Markov chain* to represent that a subsequent part of a bigram depends on its *immediate antecedent only*. This metric is coined *Markov* MMI (MMMI) by the present research.

It is anticipated that MMMI can better quantify an n-gram's *internal cohesion* by weighting all the *glue values*, i.e. associations, of pseudo-bigrams within an n-gram such that it can represent the entire n-gram's internal cohesiveness as each glue value is *at the dispersion point in a pseudo-bigram*, i.e. "*" in $[w_1] * [w_2, w_3 \cdots w_n]$. Accordingly, the mechanism of MMMI *ameliorates the bias* of the *conventional* MMI in the calculation of cohesion between words caused by differences in n-gram *lengths*. As n-gram is a special case of concgram, internal associations in concgrams of different lengths represented as pseudo-bigrams becomes also "*measurable and comparable*."

But this scheme can handle *only the n-gram*. Despite this imperfection, MMMI is still adopted as a specification of the metric to calculate MMI in the present research. How to *generalize* the mechanism of MMMI from the setting of the n-gram into that of the concgram, which can be

contiguous (i.e. n-gram e.g. “in fact”), non-contiguous (e.g. “either…or”), and hybrid (e.g. “as far as…concerned”) MWEs? This imperfection will be dealt with in promising future research. Expediently, it is *temporarily assumed* that *non-contiguity* and *hybridity* do not affect the *statistical uncertainty and association* of words within a concgram. This may be justified by the intuition that during speaking or writing an MWE—speakers or writers *consciously* concern the *completeness* of the idiomatic pattern of an MWE, e.g. “not only…but also”, that speakers or writers want to express.

Besides, Feng & Hu (2012) apply the concept *Shannon entropy* in classical information theory to the quantification of *uncertainty among words*. There is more than one version of entropy, including Shannon entropy, conditional entropy etc. But Shannon entropy, the most basic scheme, is adopted in the present research. It is the *dependent variable* in the present research’s *multiple linear regression model*, which complies with William of Ockham’s *Law of Parsimony*. According to Shorter Oxford English Dictionary, this law states that “in explaining a thing no more assumptions should be made than are necessary.”

And only MWE Length and MMMI are the *independent, predictive or explanatory variables* in the multiple linear regression modelled in the present study. The main purpose of architecting this regression model is to *quantify* whether the dependent variable Shannon Entropy *positively or negatively correlated to* the two independent variables. Accordingly, there are two null hypotheses, H_0’s, and two alternative hypotheses, H_1’s, about the regression coefficients that can be set up:

H_0: Shannon Entropy and MWE Length are not statistically *positively* correlated.

H_1: Shannon Entropy and MWE Length are statistically positively related.

H_0: Shannon Entropy and MMMI are not statistically *inversely* correlated.

H_1: Shannon Entropy and MMMI are statistically inversely related.

Translating the findings into linguistic interpretations: Shannon Entropy and MMMI are *inversely* correlated to a statistically significant extent. Does this imply that shorter MWEs tend to be more indeterminate than longer ones? But Shannon Entropy and MWE Length are *positively* correlated to a statistically significant extent. Does this imply that shorter MWEs tend to be more determinate than longer ones? These two relations appear to be contradictory, but $\log_2 p_i$ in the specification of Shannon entropy below must be *negative*, while additional variables’ probabilities are added to the cumulative total, which is *made positive by the negative sign*:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

It is implied that H increases accordingly. Thus, that Shannon Entropy increases as MWE Length increases can be *logically inferred*. Statistically, the *correlation* between Shannon Entropy and MWE Length is 0.683, which is positive and regarded as high. From another viewpoint, the

seemingly contradictory empirical relations can be diagnosed as follows: MWE Length is a very primitive variable as it is just a discrete quantitative variable with values 2, 3, 4 and 5 only. In contrast, MMMI is a continuous quantitative variable obtained by sophisticated derivation. Hence, it should be not appropriate to directly and superficially compare those two empirical relations.

Finally, it is *temporarily and expediently assumed* that non-contiguity or hybridity in an MWE does not affect the *statistical uncertainty and association of words* within a concgram-based MWE, but can this issue be left open? Take a concgram-based MWE extracted from ICLE as a concrete example: "as far as" + "my subjective point of view is" + "concerned". "as far as…concerned" is a hybrid concgram. Can "my subjective point of view is" statistically qualify as *intervening words*, which have no effect on the validity of MMMI based on Wei & Li's (2013 pp. 508) formulation? This is a major question of *statistical idiomaticity* to be addressed in prospective research.

As to *formulaic idiomaticity*, to what extent is an MWE formulaic in the sense that it consists of formulaic components, which *cannot be structurally dissected*, and non-formulaic components? From ICLE, the concgram-based MWE 'as far as' + 'my subjective point of view is' + 'concerned' is taken apart by Alex's (or Survey) Parser that Fang (2007) conveys. This parse is shown below:

```
PU CL(depend,act,indic,intr,pres,sub,unm)
 DISMK CONNEC(ge) {And}
 SUB SUBP()
  SBHD CONJUNC(subord) {as far as}
 SU NP()
  DT DTP()
   DTCE PRON(poss,sing) {my}
  NPPR AJP(attru)
   AJHD ADJ(ge) {subjective}
  NPHD N(com,sing) {point}
  NPPO PP()
   P PREP(ge) {of}
   PC NP()
    NPHD N(com,sing) {view}
 VB VP(pass,pres,montr)
  OP AUX(pass,pres) {is}
  MVB V(montr,edp) {concerned}
 PUNC PUNC(com) {,}
 PUNC PUNC(per) {.}
```

Can this parse reveal formulaic idiomaticity? "as far as" is obviously a formulaic component. But how about "as far as…concerned"? How about the non-formulaic? Future research is anticipated.

Corpora and Empirical Approaches

To highlight the distinctive features of the present research, this section is purposefully composed of three sub-sections: Corpus-based plus Corpus-driven Approaches; Quantitative Turn in Learner Corpus Research; and Overarching Bottom-up Approach. It is deliberately formed to avoid that readers would overlook these features.

Corpus-based plus Corpus-driven Approaches

Tognini-Bonelli (2001 pp. 65) succinctly defines 'corpus-based' as a qualifier which means *expounding, testing or exemplifying theories and descriptions based on corpus*, whereas Tognini-Bonelli (ibid. pp. 84) defines 'corpus-driven' as a qualifier which means using "a corpus *beyond* the selection of examples to support linguistic argument or to validate a theoretical statement". In other words, corpus is not only "a repository of examples to *back pre-existing theories or a probabilistic extension* to an already well-defined system" but also "many of the *(theoretical) statements are of a kind that are not usually accessible by any other means than the inspection of corpus evidence*".

Likewise, Biber (2012 online) insightfully states that corpus-based language research "*assumes the validity of linguistic forms and structures derived from linguistic theory*" such that research mainly aims at analyzing "the systematic patterns of variation and use for those *pre-defined* linguistic features". In contrast, corpus-driven research is relatively *more inductive* such that "the linguistic constructs themselves *emerge from analysis of a corpus*."

The present research adopts both "corpus-based" and "corpus-driven" approaches. On the one hand, it is corpus-based because a set of idiomatic multi-word expressions (MWEs) are acquired from website Smart Words (2013) and these MWEs are *'proved' to be sufficiently present in* the *International Corpus of Learner English* (ICLE). On the other hand, it is corpus-driven because the relations between Shannon entropy, Markov multivariate mutual information (MMMI), and MWE length are *measured empirically via a multiple linear regression model driven by text data in ICLE*. These terminologies will be explicated in section Research Methodology with Literature Review.

In order to be *fully corpus-driven*, it is recommended that *nonparametric regression* should be adopted in future research as Hazelton (2015) pinpoints that "in nonparametric regression we make almost no assumptions about the shape of the relationship, letting the data 'speak for themselves' in determining the estimated trend." But multiple linear regression, a kind of *parametric regression*, does have its merits. Thus, there exists *semiparametric regression*, a synthesis of parametric and

nonparametric regression, which will be expounded in sub-section Regression Modeling under section Future Research.

Quantitative Turn in Learner Corpus Research

Probabilistic or statistical corpus and computational linguistics are relatively new paradigms in modern linguistics. They are explored by Geman & Johnson (2003) and by Goldsmith (2003). Accordingly, probability can be considered highly relevant to linguistic research. It is also the foundation of modern statistics though not many statistics textbooks give sufficient coverage on probability theory.

Gries (2013) takes the lead to apply *statistical tests* to the analytics of authentic learner corpus data in the context of learner corpus research (LCR). Nearly all sub-branches of linguistics are currently evolving into "much more *empirical*, much more *rigorous*, and much more *quantitative / statistical*" enterprises. "While most, though of course not all, of the 20th century linguistics was characterized by a reliance on what some have referred to as armchair linguistics, where a linguist develops a theory and at the same time makes up the data—usually *acceptability judgments of decontextualized isolated sentences*". This circumstance has been a complete change for the better. At least many, if not all, linguistic sub-fields now routinely have empirical studies that "use experimental designs and/or sophisticated analyses of corpus data". In tandem with this trend of *increasing empiricism*, the corresponding revolution is that adopting "more *rigorous statistical analysis* of various levels of complexity, especially inferential statistics, is becoming a major element of linguistic analysis."

A new but peer-reviewed *International Journal of Learner Corpus Research* was launched in 2015. Despite its short history, this flagship signifies a milestone in the prospect of conducting LCR. Gries & Deshors' (2015d) paper in this journal particularly merits a mention as it adopts *multi-level regression modeling* and highly recommends using such rigorous methodological approaches in LCR. This is encouraging as the present research also harnesses a regression model but of a different kind, multiple linear regression, to marshal the analytics of seemingly unrelated raw data on corpus.

Even further, in a forthcoming paper Paquot & Plonsky (2017) delve into the application of quantitative analysis to empirical assessment of the "study quality" in LCR. They highlight that the *insufficiency in* "*statistical literacy*" should be addressed. Their study targets at conducting "the first empirical assessment of quantitative research methods and study quality in learner corpus research". They methodically evaluate existing "quantitative primary studies referenced in the Learner Corpus Bibliography (LCB)", representative of LCR and compiled by the Learner Corpus Association. LCB is composed of 1,276 references since the project commenced. Insightful results

reveal "several systematic strengths as well as many flaws", e.g. the *lack of research questions, partial and "inconsistent reporting practices* (e.g. means without standard deviations)", and the "*lack of statistical literacy* (i.e. LCR studies generally *over rely on tests of statistical significance, without reporting effect sizes, rare checking for or reporting statistical assumptions, rare use of multivariate analyses*)." Still, LCR is slowly undergoing a methodological reform: progressive enhancements are "clearly noted and there are signs that like [those] in other related disciplines".

As readers can see, the present research adopts multiple linear regression, a kind of parametric modeling. But as per Hazelton (2015), "*nonparametric regression* is a methodology for describing the trend between a response variable and one or more predictors. This approach *differs from classical regression models* in that it does not rely on strong assumptions regarding the shape of the relationship between the variables." Instead, the data can "speak for themselves in determining the form of the fitted regression functions." Although nonparametric regression has its advantages, the merits of parametric regression cannot be completely neglected. Therefore, another kind of methodology which is a synthesis of the above two regression emerges—semiparametric regression. Parametric, nonparametric, and semiparametric regression will be explicated in sub-section Regression Modeling under section Future Research.

Overarching Bottom-up Approach

It is deliberately to single out this theme as an independent sub-section and put it near the beginning of the whole research to avoid the possible doubt about why the research does not start with research questions and hypotheses. The overarching approach is to generate them in a *bottom-up fashion*. This implies that there seems to be *no firm pre-existing theoretical stand* on the thematic relation between uncertainty, represented by Shannon entropy, and associativity, represented by MMI, within concgram-based multi-word expressions (MWEs) before the present research commences to *statisticize the raw and authentic data* from ICLE into metric Shannon entropy and metric Markov multivariate mutual information (MMMI), a term coined in the present research. These metrics are to be expounded on in section Research Methodology with Literature Review. After such statistics are available, the present research then ascertains the above relation between uncertainty and associativity and that between uncertainty and MWE length by fitting a *multiple linear regression model* to the statistics. Research questions and hypotheses can then be formulated. And the ultimate outcomes are empirically-determined relations to be covered in section Empirical Analytics and Discussion. But first go over the linguistic phenomena.

Linguistic Phenomena with Literature Review

This section highlights what concepts of the linguistic phenomena the present research addresses and relates these concepts to three domains—learner language, multi-word expressions (MWEs), and concgram-based MWEs. In short, learner language means the usage of L2s performed by their learners who have the same L1 or different L1s; MWE means a phrase with two or more words that are *contiguous* (i.e. a property of the n-gram) and *non-contiguous*, or with three or more words that are *hybrid* (i.e. contiguous plus non-contiguous e.g. "not only…but also"); concgram, a new paradigm in MWE research, represents the specification of *phraseological constituency* (e.g. "AB, A*B, A**B, etc.") and of *phraseological positional variation* (e.g. "AB, BA, B*A, etc.").

Learner Language

Studied by Gilquin & Granger (2015), learner language, L2s used by learners of the same L1 or different L1s, is the target of investigation in the arena learner corpus research (LCR), which encompasses all corpus-based or corpus-driven works related to learner corpora, a type of corpora that is dedicated to depositing second languages of learners. Among such corpora, the International Corpus of Learner English (ICLE) is adopted as the authentic source of learner language in the present research.

Learner Corpus Research vs Second Language Acquisition

It is Gilquin & Granger (2015) who investigate learner language. "While a large range of language varieties had been explored by corpus linguists from the emergence of the field, it was only" three decades ago (near the end of the 1980s) "that corpus linguists began to" interest themselves in learner language by *designing and building learner corpora*, i.e. electronic collections of writing or transcribed speech produced by foreign or second language (i.e. L2) learners. This new research paradigm is commonly known as learner corpus research (LCR), having broken into "a field which until then had been the sole remit of second language acquisition (SLA)".

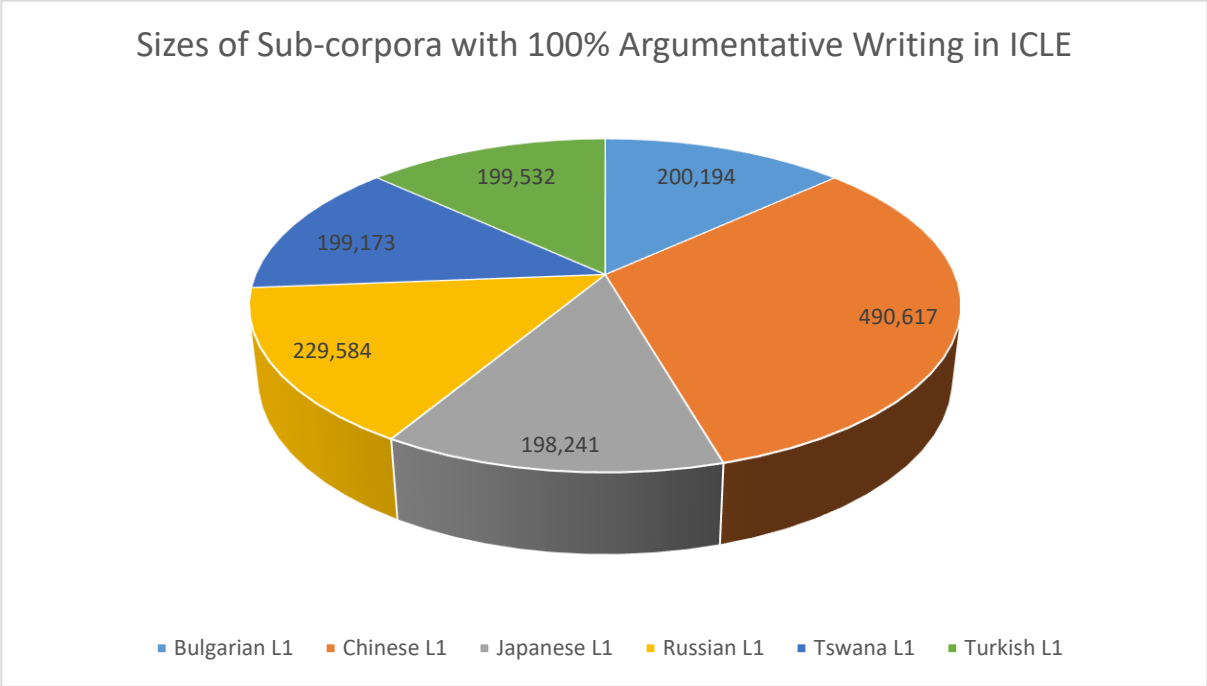
These two fields, LCR and SLA, share the same research interest, "i.e. learner language, but they differ markedly in their objectives and methods of analysis." One of the major discrepancies is that SLA studies focus on *competence* whereas LCR studies on *performance*. LCR aims at describing how L2 learners use language in actual production of the L2 by deploying "the *tools and techniques of corpus linguistics*", which are of a high degree of *automation* that enables "the study of whole *learner populations*". In contrast, traditional SLA studies use the *more manual analytical methods* "which are better suited for the investigation of a small number of *individual learners*".

By contrasting LCR with SLA, Gilquin & Granger (ibid.) can bring out the biggest dissimilarity between these two fields. Indeed, such a comparison immediately enlightens readers on the fields' major differences. This particularly informative book chapter also inspires the present research as it can highlight the major dissimilarities that guide the present research to clearly focus on what LCR is and is not.

The Design of International Corpus of Learner English & its Derivative

Knowing the *design* of a corpus seems unnecessary and is even deemed a waste of time, but it is indeed crucial to the understanding of how linguists can assess whether their *research inquiries* can be informed by processing the data on a *particular corpus*. The International Corpus of Learner English (ICLE) v.2 was launched by Granger, Dagneaux, Meunier & Paquot (Eds.) (2009), offering authentic data of *naturally occurring* learner English for the purpose of undertaking empirical investigation. ICLE v.2 is an expanded version that has *16 sub-corpora of texts* written by learners of English from 16 different countries of 16 L1s (4.5 million words in total).

As to the characteristics of text data in ICLE v.2 that have impacts on how linguists interpret the findings based on this corpus, Granger, Dagneaux, Meunier & Paquot (ibid pp. 7) pinpoint that most of the essays in it belong to *untimed essay written at home*: "untimed" (62%), "not written under exam settings" (61%), and "written with reference aids" (48%). The implication could be that the learner language on the corpus might be *closer to that of native tongue than expected* because of the availability of *time and aids* to write the essays. Here is a pie chart showing the number of word tokens in its sub-corpora of essays composed by learners of English speaking 6 different L1s.



In the present research, only these 6 sub-corpora (*1.12 million words in total*) rather than all 16 sub-corpora in ICLE are compared. An *aggregate* sub-corpus of these 6 sub-corpora is constructed by merging the 6 individual sub-corpora. Why only these 6 sub-corpora? It is because the learner English essays are *100% argumentative across the board* such that the implicit impact of the *factor 'genre' is consequently controlled*. The distribution of the number of word tokens over the 6 L1s, i.e. Bulgarian, Chinese, Japanese, Russian, Tswana, and Turkish, is illustrated on the pie chart above.

All the empirical findings in the present research are based on the individual statistics of the 6 shortlisted sub-corpora in ICLE and the consolidated findings of the aggregate sub-corpus of these sub-corpora: *L1-based and general snapshots* of those *contiguous*, *non-contiguous*, or *hybrid* concgram-based MWEs in L2 (i.e. English as a second language), prespecified according to website Smart Words (2013), and supplemented by the present author's intuition. This seeming unauthenticity of the MWEs on the website is not a concern as they are *checked against the authentic data* in ICLE. In this regard, the present research is *corpus-based*; but it is shown to be *corpus-driven* as well (see sub-section Multiple Linear Regression in section Research Methodology with Literature Review).

Learner Corpus and Studying Multi-Word Expressions

Paquot & Granger (2012) examine the *formulaic language* on learner corpus, and the present study's research topic is in line with their theme. Formulaic language is a pivotal phenomenon not only in corpus linguistic research in general but also in learner corpus research (LCR) in particular. And for consistency of terminology, "multi-word expressions" (MWEs) is used hereafter.

Corpus linguistic techniques are a tremendously powerful means of exploring MWEs; in fact, they were under the spotlight from the very early days of LCR. Firstly, "the focus is on the types of learner corpus data investigated and the most popular method used to *analyze* them". Secondly, research on *formulaic* MWEs "describes the types of word sequences analyzed in learner corpora and the methodologies used to *extract* them". Finally, Paquot & Granger's (ibid.) paper summarizes some of the key findings of LCR studies on "distinguishing between *co-occurrence* and *recurrence*". And particular emphasis is also put on the relationship between learners' use of formulaic MWEs in L2 and the *transfer* of influence from learners' L1s.

Further, Ebeling & Hasselgård (2015 pp. 208) suggest that some may claim that identifying MWEs in corpora has two distinctive ways—the *bottom-up* and the *top-down*. "While the former is associated with *corpora* and the *frequency-based approach to* phraseology, the latter is firmly anchored in the *phraseological* approach. Conventionally, frequency would not interest the phraseological approach, which thrives on "prototypical (low-frequency) *idioms* such as kick the

bucket". But the situation *has changed in recent years* such that the investigation of "*predefined sequences* or categories of words" has increasingly relied on the use of *corpus data and frequency-based metrics*. It seems that "reconciling the frequency-based and the phraseological approach is a step in the right direction."

Another aspect of phraseology is *formulaic language*—one kind of MWEs in the context of second language acquisition (SLA) examined by Ellis, Simpson-Vlach, Römer, O'Donnell & Wulff (2015 pp. 358). Research "in *psycholinguistics, corpus linguistics* and *cognitive linguistics*" reveals that "language users have rich knowledge of the frequencies of forms and of their sequential dependencies in their native language". Indeed, language processing is responsive to the "*sequential probabilities*" of linguistic constituents, from micro to macro levels i.e. from phonemes to phrases, "in comprehension as well as in fluency and idiomaticity of speech production".

Ellis, Simpson-Vlach, Römer, O'Donnell & Wulff (ibid. pp. 359) suggest that such sensitivity to *sequentiality in language processing* authenticates the claim that learners have "*implicit knowledge of memorized sequences of language*", and this competence is "the basis for *linguistic systematicity and creativity*". These two seem mutually exclusive. But it is the last decade that there is tremendous confirmatory evidence of "native and L2 users' implicit knowledge of *linguistic constructions and their probabilities of usage*."

Multi-word Expressions (MWEs)

This section seemingly comprises a disparate and loosely-coupled array of concepts. But, in fact, it is concept-centric—juxtaposes them, which might be overlooked by readers. The conception MWE is exactly the context in which another conception concgram resides. And concgram, which encompasses *phraseological constituency* (e.g. "AB, A*B, A**B, etc.") and *phraseological positional variation* (e.g. "AB, BA, B*A, etc."), is a transition from traditional conception n-gram and newer conception skipgram, which depicts non-contiguous word association.

Idiomaticity vs Compositionality: Continuum?

Baldwin & Kim (2010 pp. 269) provide an insight into idiomaticity in the context of MWEs. Idiomaticity denotes "*markedness* or deviation from the basic properties of the *component lexemes*, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels." It is *statistical idiomaticity* that is particularly relevant to the present research. Such "idiomaticity occurs when a *particular combination* of words occurs with *markedly high frequency*, relative to the component words or alternative phrasings of the same expression".

Seemingly an all-or-nothing notion contradictory to idiomaticity is *compositionality*. As per en.oxforddictionaries.com, "all-or-nothing" means "having no middle position or compromise available". Baldwin & Kim (ibid. pp. 269) define compositionality as "the degree to which the features of the parts of an MWE *combine to predict* the features of *the whole*." Is it necessary that there cannot be a *co-existence* of a certain *degree of idiomaticity* and a certain *degree of compositionality*? Martinez & Murphy (2011) state that evidence reveals "a significant number of relatively opaque expressions occur frequently in texts in English. One commonly cited estimate prepared by Erman & Warren (2000) is that somewhat more than one-half (55%) of any text will consist of formulaic language (p. 50)."

Conklin & Schmitt (2012 pp. 45) elucidate that "…we store representations of individual words in our mental lexicon." There is a converging consensus that human mind's "lexicon also contains formulaic language (*How are you? kick the bucket*). In fact, there are compelling reasons to think that the brain represents formulaic sequences in long-term memory, bypassing the need to compose them online through word selection and grammatical sequencing in capacity-limited working memory."

Lexico-grammatical Units

Abu-Ssaydeh (2006 pp. 351, 353) succinctly accentuates that a multi-word unit (MWU) is a "*lexico-grammatical unit*", which means that the lexicon of a language is entwined with the grammar of that language. For consistency of terminology, the term MWUs is hereafter called MWEs. In other words, MWEs obscure the conventional distinction linguists incline to make between the two hierarchies lexis and syntax. Elaborately, MWEs embrace "*collocations, idioms, binomials and trinomials, complex adverbs, prepositional phrases and formulaic expressions...*". Further, Abu-Ssaydeh (ibid.) adumbrates that *empirical psychological evidence* reveals how our mental lexicon retains MWEs as aggregate modules harnessed by the human mind to fulfill "discoursal, stylistic and pragmatic purposes".

Collocation Tendency

Zooming in on an issue in linguistics in general—collocation, Kumova Metin & Karaoğlan (2011) aim at using quantitative analysis to measure "*collocation tendency of words*". Their design is that a "collocation tendency method is tested on a base data set extracted by some *statistical collocation extraction techniques* (frequency of occurrence, point-wise mutual information, t-test, chi-square technique) and is evaluated by *precision and recall measures*," of which definitions are explicated on webpage en.wikipedia.org/wiki/Precision_and_recall: "In simple terms, high precision means

that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results."

Among those techniques, it is mutual information (MI) that is particularly relevant to the present research. Specifically, it is Markov multivariate MI (MMMI) rather than Shannon's original bivariate, or point-wise, MI is adopted. How this bivariate specification is transformed into MMMI is proposed by Wei & Li (2013). MMMI is a term coined in the present research, given that the concepts behind MMMI are from the da Silva & Lopes (1999) and Wei & Li (ibid.) but the term is not.

Formulaic Fixedness vs Internal Variability

Nissim & Zaninello (2013) proclaim that "the issue of internal variability" of multiword expressions (MWEs) is of paramount importance to the *identification and extraction of them in running text*. They then impart corpus-based analytics of Italian MWEs, aiming to define a mechanism "for modeling internal variation [and] exploiting *frequency* and *part-of-speech (POS) information*." But they highlight that "since a *search for fixed forms* suffers from low *recall*, while an *unconstrained flexible search* for lemmas yields a loss in *precision*", they advance a machinery that aims at "maximizing precision in the identification of MWEs within a flexible search". Their approach thrives on "the idea that *internal variability* can be modelled via the novel introduction of *variation patterns, which work over POS patterns*, and can be used as working tools for controlling precision". They also benchmark "the performance of variation patterns to that of *association measures*", and seek out "the possibility of using variation patterns in MWE *extraction* in addition to *identification*."

Biber (2009 pp. 275) exploits a *corpus-driven* methodology to discriminate between "the most common multi-word patterns" in interlocution and those in academic writing, and to inspect the asymmetric types of patterns in these two registers, i.e. conversation and writing. Biber (ibid.) first conducts surveys to uncover the distinguishableness of the methodological features of corpus-driven research in linguistics and then juxtaposes the linguistic attributes of two kinds of "multi-word sequences": "*multi-word lexical collocations*" (i.e. "combinations of content words") versus "*multi-word formulaic sequences*" (i.e. "incorporating both function words and content words"). The all-important precedence is to empirically examine the 'patterns' exhibited in multi-word formulaic sequences. The main finding is that such patterns typical of speaking rudimentarily contrast with those typical of writing: the former tend to be "*fixed sequences* (including both function words and content words)", whereas the latter are "*formulaic frames* (consisting of *invariable function words* with an *intervening variable slot* that is filled by content words)."

Concgram-based MWEs

Concgrams can be contiguous, non-contiguous, or hybrid (i.e. contiguous plus non-contiguous) multi-word expressions. The present research would like to pinpoint that concgrams are the outputs of phraseological search engine ConcGram©, which can support both, rather than either one, of *corpus-based and corpus-driven research* (see sub-section Corpus-based plus Corpus-driven Approaches) into MWEs though the literature on concgramming states that it is construed as corpus-driven.

Phraseological Constituency and Positional Variation

Comprehensively, Cheng, Greaves & Warren (2006) and Greaves & Warren (2010 pp.218) expound on the conception concgram, which addresses the deficiency of classical MWE constructs, including the n-gram and the skipgram, by picking out *phraseological constituency* (e.g. "AB, A*B, A**B, etc.") and *phraseological positional variation* (e.g. "AB, BA, B*A, etc."). Then, *concgramming* means the process in which corpus-based and corpus-driven software ConcGram© mines all the co-occurrences of at least two but at most five words irrespective of variations in constituency and positions of words. And Cheng, Greaves, Sinclair & Warren (2008) contend that studying the concgram can facilitate "taking us closer to more fully appreciating and understanding the *Idiom Principle*".

The raw frequencies of non-contiguous or hybrid concgram-based MWEs selected in the present research need to be semi-automatically counted as there are *intervening words, a label which we tentatively coin*. For example, in the following sequence of words "And **as far as** my subjective point of view is **concerned**," its idiomatic pattern is a 4-word hybrid concgram "as far as…concerned", while its *presumptive* intervening words are "my subjective point of view is". As Biber (2009 pp. 275) advances, his research also asserts that a non-contiguous or hybrid concgram contains both an *idiomatic pattern*, which we call concgram, and one or more *slots of intervening words*.

Research Methodology with Literature Review

As mentioned earlier in section Corpora and Empirical Approaches, an *overarching bottom-up approach* to empirical research is adopted in the present study. This section first elaborates mutual information as an association measure, then Shannon entropy as an uncertainty measure, and what multiple linear regression is. Afterward, spreadsheeting plus statistical computing with R, as two complementary tools, are expounded in tandem. Next, how a computational approach is applied to concgramming is enunciated. With all these topics covered, research questions and hypotheses can then be set out.

Mutual Information as an Association Measure

Wei & Li (2013 pp. 507) pinpoint that "*raw frequency* cannot be a good indicator of saliency of association of word sequences" because it can be impacted on by *corpus size*. This leads to the need of a metric that is free from this issue. And such a metric is *mutual information* (MI).

Shannon's Mutual Information

Shannon entropy, explicated by Goldsmith (2003 pp. 23-24), and mutual information, expounded on by Goldsmith (2003 pp. 24-25), are two intertwined core concepts in *classical information theory*. Of the conceptions in this theory, bivariate mutual information (BMI) has been applied by Church & Hanks (1990) to *measuring word association norms* in linguistics and lexicography. Application of this theory to the study of phraseology is conducted by Gray and Biber (2015). Such an *association measure* BMI is depicted by Church and Hanks (1990, pp. 23), as cited below:

$$I(x, y) \equiv \log_2 \frac{P(x, y)}{P(x)P(y)}$$

Here $I(x, y)$ is BMI, a *specific instance* rather than an *expected value* that will be mentioned later. In this formulation, $P(x)$ and $P(y)$ are the probabilities of two random variables, e.g. words. As per Wikipedia, en.wikipedia.org/wiki/Binary_logarithm, $\log_2 n$ stands for *binary logarithm* in mathematics and is the *power* to which the *base* number 2 must be raised to obtain the value of variable n. One key point on this webpage is that binary logarithm can be utilized to ascertain the number of *bits* for encoding a digital message in classical information theory.

Church and Hanks (1990, pp. 23) continue to informally unpack that BMI can be interpreted as a *ratio* that compares the *joint probability* of the occurrence of both x and y with the *independent probabilities* of the independent occurrences of x and y. A further interpretation is that if the occurrences of x and y are *genuinely associated* with each other, their joint probability $P(x, y)$ is

much greater than the product of their independent probabilities $P(x)$ and $P(y)$ so $I(x,y) \gg 0$, which means "much greater than zero". But if *no genuine*, no matter positive or negative, *association* exists between x and y, the numerator $P(x,y)$ is close to the denominator $P(x)P(y)$ in the ratio. Mathematically, $P(x,y) \approx P(x)P(y)$, hence $I(x,y) \approx 0$, which means "close to zero". In contrast, if *genuine dissociation* exists between the occurrences of x and y, $P(x,y)$ is much less than $P(x)P(y)$ such that $I(x,y) \ll 0$, which means "much less than zero".

Of course, the probabilities $P(x)$ and $P(y)$ must be estimated to obtain numerical values of $I(x,y)$. These two probabilities can be calculated by using software or programming to count the observations or instances of x and y, denoted as $f(x)$ and $f(y)$, on a corpus, and then $f(x)$ and $f(y)$ are *normalized by the population size*, i.e. the size of the corpus, usually denoted as N. Besides, the joint probability of the co-occurrences of x and y, i.e. $P(x,y)$, can be estimated by numerating the instances of x and y occurring in tandem, i.e. $f(x,y)$, which is then normalized by N.

Prior to the explanation of a method for generating *multivariate mutual information* (MMI), the present research adumbrates how Van de Cruys (2011) depicts the interpretation of BMI, which means a probabilistic measure that quantifies the *amount of information* contained in one random variable (RV) about another random variable. With more information about another random variable, the *uncertainty* about that variable is then *reduced due to the information of the other*.

Here is the formula for the *expected value* of BMI:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Notationally,

1) $\sum$ means summation;
2) $\in$ means 'is an element of' or 'is an instance of';
3) Both X and Y represent two random variables, of which *instances* are x and y; and
4) $\log_2 z$ stands for the logarithm at base 2 (i.e. binary logarithm) of the number z.

In principle, the expected value can be construed as a *weighted average* of all instances $\log_2 \frac{p(x,y)}{p(x)p(y)}$, on the left of which is probability $p(x,y)$—the *weight*.

This formula says that BMI is an association measure about specific instances of the two RVs X and Y. Put it differently, this measures the ratio between the probability of co-occurrence of these two RVs, i.e. their *joint probability distribution* $p(x,y)$, and the probability of co-occurrence of them, i.e. the *marginal probability distributions* $p(x)$ and $p(y)$ of these two RVs, assuming the independence between these two RVs. Importantly, the formulation of BMI in the equation above specifies that the expected value is derived from *all possible instances of occurrence* of the two

RVs X and Y. And $I(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$ can be positive or negative or close to zero, but its *expected value* $I(X;Y)$ over all stochastic events, i.e. the occurrences of words, is *always non-negative*.

Pseudo-bigram Transformation and Markov Chain

Depending on how it is presented, a Markov chain, an admired kind of *discrete stochastic process*, can be quite intimidating if it is articulated rigorously. But the *fundamentals* of probability and Markov chain, which always go hand in hand, are explained by Ghahramani (2016) in a digestible but not simplistic manner. This special kind of Markov process is used by Wei & Li's (2013) journal article which particularly belongs to quantitative and corpus linguistics.

For empirical modeling of conventional n-gram (a special case of concgram, i.e. contiguous concgram), there is an alternative specification which Wei & Li (2013) devise and the present research coins—*Markov multivariate mutual information* (MMMI), which combines the construct *pseudo-bigram* with *first-order Markov chain or model*. This construct pseudo-bigram is first asserted by da Silva & Lopes (1999), and is adopted by Wei & Li (2013 pp. 512) to represent an n-gram as a set of recursively dissected pseudo-bigrams, e.g. $[w_1] * [w_2, w_3 \cdots w_n]$, $[w_1, w_2] * [w_3 \cdots w_n]$ and so on.

Wei & Li (2013 pp. 512) state that "…every n-gram may be thought of as a pseudo-bigram in terms of having one *dispersion point* located between *a left and a right part* of the n-gram: $w_1 \ldots w_i$ and $w_{i+1} \ldots w_n$ $(n \geq 2)$." They continue to extend "the theory of *pseudo-bigram transformation*" on the conceptual basis of first-order Markov Model (MM). First, they explicate a first-order MM, and then an n-gram's dispersion points.

As to a first-order MM (ibid. pp. 513), it is used to describe a *sequence* of random variables, $W = [w_1, w_2, \cdots, w_n]$, that are *not independent* of each other, i.e. "the value of each variable relies on previous elements in the sequence". Nevertheless, it is assumed that "only the prior local context", i.e. the previous $n-1$ words, "affects the occurrence of the next word". In theory, "the larger value of $n-1$, i.e. the *more previous words* that are included for predicting the next upcoming word, the *more accurate* an MM will be." But "the larger the value of $n-1$ is, the *more parameters* need to be estimated, and the *more intractable* the calculation will be". Thus, in linguistic studies, researchers usually use a first-order MM, which "serves as the conceptual framework for transforming a *multi-word sequence* of any lengths into a *pseudo-bigram*". Multi-word sequences are conceptually equivalent to multi-word expressions (MWEs).

As for an n-gram's *dispersion points* (ibid. pp.513), a dispersion point is conceptually the space that "may locate between the positions of the constituent words of an n-gram". After that point and before it, several other words may appear, showing a kind of *dispersion tendency* at the point. Take "a recurrent bigram *United States*" as an example. Despite the probability for these two words to occur together is high, however, "it is still possible to find instances where the two words do not appear together, as in the cases of *United Kingdom … United Europe*, *United Airlines …*" This implies "a kind of dispersion tendency for different words in each inter-word space in the bigram" exists. "For any arbitrary n-gram $w_1, w_2, w_3, \cdots, w_n$, there are $n-1$ dispersion points". This n-gram is recursively dissected into a set of pseudo-bigrams. The sign $*$ below denotes a dispersion point, which is moving rightward:

$$[w_1] * [w_2, w_3, \cdots, w_n]$$
$$[w_1, w_2] * [w_3, \dots, w_n]$$
$$\vdots$$
$$[w_1, \dots, w_{n-2}, w_{n-1}] * [w_n]$$

Markov Multivariate Mutual Information (MMMI)

The present research proceeds to obtain MMI numerically derived from the tokens in a concgram-based MWE for proving that a concgram-based MWE is *statistically significantly* co-occurring, i.e. contiguous (e.g. for instance), non-contiguous (e.g. so…that) or hybrid (e.g. not only…but also). To what extent are the *tokens* that constitute a concgram-based MWE *associated with each other*?

Markov multivariate mutual information (MMMI), coined by the present research, is based on a mathematical specification extended from the original bivariate mutual information (BMI), which synthesizes *pseudo-bigram* with *first-order Markov chain* due to Wei & Li (2013). MMMI is expected to be a reliable and meaningful metric that helps linguists to objectively and efficiently assess whether a concgram-based MWE is a statistically cohesive unit of linguistic patterns, of which the internal relations may not be analyzable in terms of conventional linguistic theories.

It is necessary to first understand what a Markov chain is because Wei & Li's (ibid.) new computing method for specifying multivariate mutual information requires the use of such a *discrete-time stochastic or random process—Markov chain*, which is expounded on by Ghahramani (2016 pp. 494-496): "A stochastic process $\{X_n: n = 0,1, \cdots\}$ with a finite or countably infinite *state space* S is said to be a Markov chain, if for all i and j, $i_0, \dots i_{n-1} \in S$, and $n = 0,1,2 \dots$" such that the following specification holds:

$$P(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i).$$

Verbally speaking, n is the *unit of time* and this mathematical expression means that given the state at *present* $(X_n = i_n)$, its *future* state $(X_{n+1} = i_{n+1})$ is independent of the *history* or past states $(X_{n-1}, \cdots, X_1, X_0)$. In addition, a Markov chain belonging to state space S is regarded as having stationary transition probabilities if, for all $i, j \in S$, the *transition probability* of a one-step movement from state i to state j is *independent of the time* at which the transition will occur. That is, a Markov chain is *stationary* if $P(X_{n+1} = j \mid X_n = i)$ does not depend on n, the unit of time.

Mathematically speaking,

For $i, j \in S$, if

$$p_{ij} = P(X_{n+1} = j \mid X_n = i);$$

then the *transition probability matrix* is:

$$P = \begin{bmatrix} p_{00} & \cdots & p_{0j} \\ \vdots & \ddots & \vdots \\ p_{i0} & \cdots & p_{ij} \end{bmatrix}$$

Its interpretation is that the *sum* of the elements (i.e. probabilities) of *each row* of P is 1. That is, the following equation holds:

$$\sum_{j=0}^{\infty} p_{ij} = 1$$

An intuitive example is used to illustrate the abstract concepts above: A working traffic light will be out of order tomorrow with probability 0.07, and an out-of-order traffic light will be working again tomorrow with probability 0.88. Let $\{X_n : n = 0, 1, \ldots\}$ be a Markov chain with state space $\{0, 1\}$, and transition probability matrix is as follows:

$$P = \begin{bmatrix} 0.12 & 0.88 \\ 0.07 & 0.93 \end{bmatrix}$$

Clearly, the sum of probabilities on each row is equal to 1, i.e. 0.12+0.88=1 and 0.07+0.93=1.

To address the deficiency of existing method using "an *arithmetic average* of the products determined by each dispersion point along the *n*-gram", Wei & Li (2013 pp. 516-518) propose an ingenious "normalizing algorithm of a *probability-weighted average* for calculating an n-gram's internal association". In this research, the steps involved in the mathematical derivation of and examples of synthesizing constructs pseudo-bigram and Markov chain are shown below:

1. Calculate the *expected joint probability* of each dispersion point by multiplying the mass probability of each part of the *pseudo-bigram*, i.e. $P(W_1, \ldots, W_i) \cdot P(W_{i+1}, \ldots, W_n)$; for example, "as well as" is a frequent concgram-based MWE in ICLE:

   $E_1 = E_{(as*well\ as)} = P_{as} \times P_{well\ as} = 0.0090 \times 0.0002 = 2.20 \times 10^{-6}$

   $E_2 = E_{(as\ well*as)} = P_{as\ well} \times P_{as} = 0.0008 \times 0.0004 = 3.38 \times 10^{-7}$

2. Take the *probability-weighted average* of all the *expected joint probabilities*, denoted by WAP, where the weighting depends on the probability of each value; for instance,

$$WAP_{(as\ well\ as)} = \sum_{i=1}^{i=n-1} P_{[P(W_1 \dots W_i)\cdot P(W_{i+1} \dots W_n)]} \cdot [P(W_1 \dots W_i) \cdot P(W_{i+1} \dots W_n)$$

$$= \sum_{i=1}^{i=3-1=2} P(E_i) \cdot E_i = P(E_1) \cdot E_1 + P(E_2) \cdot E_2$$

$$= \frac{E_1}{E_1 + E_2} \times E_1 + \frac{E_2}{E_1 + E_2} \times E_2$$

$$= \frac{2.20 \times 10^{-6}}{2.20 \times 10^{-6} + 3.38 \times 10^{-7}} \times 2.20 \times 10^{-6}$$

$$+ \frac{3.38 \times 10^{-7}}{2.20 \times 10^{-6} + 3.38 \times 10^{-7}} \times 3.38 \times 10^{-7}$$

3. Take *binary logarithm* of the *quotient* of the expected joint probability, $P(W_1, W_2, \dots, W_n)$, and the weighted average of the expected joint probability, i.e. WAP, thereby obtaining the *weighted glue value* or MMMI for the n-gram as a whole.

$$MMMI(W_1, W_2, \dots, W_n) = \log_2 \left[\frac{P(W_1, W_2, \dots, W_n)}{WAP}\right]$$

$$MMMI_{(as\ well\ as)} = \log_2 \left[\frac{P_{(as\ well\ as)}}{WAP}\right]$$

$$= \log_2 \left(\frac{2.81 \times 10^{-4}}{1.95 \times 10^{-6}}\right)$$

$$\approx 7.1701$$

Shannon Entropy as an Uncertainty Measure

Why information theory is statistical is because it is mainly based on the concepts in probability theory. In the present research, why *Shannon entropy*, different from joint entropy and conditional entropy, $H = \log_2 n$ (i.e. binary logarithm or logarithm at base 2) is adopted? According to Feng & Hu (2012, pp. 344-346) [in Chinese, translated into English below], this can be accounted for by the following logical reasoning and mathematics: Before Shannon, Ralph Hartley already put forward the use of $\log_2 n$ to measure the entropy of a stochastic experiment with n possible outcomes. Using this specification is reasonable because:

1) If the number of *possible outcomes* becomes larger, the *uncertainty* of such an experiment naturally becomes larger and the entropy becomes larger accordingly.
2) If two stochastic experiments, each with n probable outcomes, are conducted concurrently, there will be n^2 possible outcomes and then the entropy will be $\log_2 n^2$. Amazingly, one of

the *laws of logarithm* is that the entropy can be transformed into $2 \log_2 n$. This projection is consistent with Hartley's view that the quantity of uncertainty will *double*.

3) Similarly, if two stochastic experiments are simultaneously conducted but one has m outcomes whereas another has n outcomes, there will be $m \cdot n$ outcomes. Then, the entropy will become $\log_2 mn$. Alternatively, one of the *laws of logarithm* states that this logarithmic expression can be transformed into the sum $\log_2 m + \log_2 n$. The uncertainty will increase.

This specification, hypothesized by Shannon and popularized by Norbert Weaver through their co-authorship, assumes that the stochastic events (in linguistics, occurrences of word token) under consideration are *equiprobable*. As per Jurafsky & Martin (2009 pp. 114-116), if such events X are not equiprobable, the above specification needs to be *generalized* as follows:

$$H(X) = -\sum_{i=1}^{n} p_i(x) \log_2 p_i(x)$$

$$= -\sum_{x \in X} p(x) \log_2 p(x)$$

This formulation can be interpreted as the *weighted sum* of each $\log_2 p_i$ averaged by probability p_i from the 1st to the nth stochastic experiment. Such an average, not arithmetic average, is known as an *expected value* in the terminology of probability theory.

Jurafsky & Martin (ibid.) contend that "most of what we will use entropy for involves *sequences*. For a grammar, e.g., we will be computing the entropy of some sequence of words" as follows:

$$W = \{w_1, w_2, w_3, \cdots, w_n\}$$

They add that "for example, we can compute the entropy of a random variable that ranges over all finite sequences of words of length n in some language L as follows:"

$$H(w_1, w_2, w_3, \cdots, w_n) = -\sum_{W \in L} p(w_1, w_2, w_3, \cdots, w_n) \log_2 p(w_1, w_2, w_3, \cdots, w_n)$$

Ghahramani (2016 pp. 494-496) also points out that "a stochastic process is said to be *stationary* if the probabilities it assigns to a sequence are *invariant with respect to shifts in the time index*." In other words, "the probability distribution for words at time t is the same as the probability distribution at time $t + 1$." Markov chains, "and hence n-grams, are stationary." For instance, "in a bigram, P_i is dependent only on P_{i-1}. Accordingly, if we shift our time index by x, P_{i+x} is still dependent on P_{i+x-1}." But Jurafsky & Martin (ibid.) single out a phenomenon that "*natural language is not stationary*" in that "the probability of upcoming words can be dependent on events that were arbitrarily distant and time dependent." Hence, a Markov chain—a probability model—"only gives an *approximation* to the correct *distributions and entropies* of natural language."

Multiple Linear Regression

The present research conducts a corpus-driven approach to quantitative analysis for empirical multiple linear regression, which is introduced in this section and of which empirical findings will be detailed in section Empirical Analytics and Discussion.

Empirical Modeling

Gries (2015a) illuminates learner corpus research (LCR) through the lens of statistical techniques, particularly *inferential statistics*, e.g. *multifactorial regression modeling*. Indeed, Gries considers regression modeling "still very much underutilized" in LCR. Such regression modeling mentioned above offers numerous merits: 1) it allows us to incorporate *multiple predictors* into an analysis; 2) such predictors facilitate scrutiny of *interactions* between variables, i.e. testing whether one variable *impacts* how another variable is correlated with the dependent variable; also, *non-linear effects* can be explored; 3) regression modeling provides a *unified framework* to understand many *seemingly unrelated tests*. For instance, instead of trying to learn many mono-factorial tests (e.g. chi-square tests, t-tests, Pearson correlations, U-tests) and then regression modeling separately, it is useful to understand that mono-factorial tests can often be seen as the simplest possible cases of a mono-factorial regression; 4) while regression modeling is typically used for *hypothesis-testing*, there are extensions that allow the researcher to also perform *data exploration*; 5) regression generates *predictions (confidence intervals)* of how a response will behave, allowing for seamless integration of results from different studies of whatever type (observational, experimental, simulations, etc.).

Mathematical Specification

Formally or mathematically speaking, an empirical model of the 'smallest' (i.e. with only two independent or explanatory variables) multiple linear regression model is adapted from Chatterjee & Simonoff's (2013 pp. 4-6) 'general' model as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

where each of the β's is known as a *coefficient or unknown parameter*, and ε_i is known as a *random error term or residual*. Both coefficients and random error terms need to be *empirically estimated* from raw data by using *statistical computing language R* to process text data in ICLE. The computing procedure with R can be very straightforward, involving only several lines of codes. The most commonly used estimation approach is known as *ordinary least squares* (OLS), which means the estimates are trialed by a computer program to minimize the following mathematical expression:

$$\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})]^2$$

Chatterjee & Simonoff (2013 pp. 7) also add that the multiple linear regression model above "can be written compactly using matrix notation" (or known as linear algebra):

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

The present research would not delve into the mathematics further since all these algebraic manipulations are seamlessly done by R. The above exposition should be enough for readers to get the gist of multiple linear regression in matrix notation.

Assumptions about Multiple Linear Regression

What really matters is that there are assumptions behind using an *OLS-based* linear regression model. Many books on statistics cover these assumptions, but Levshina (2015 pp. 155-169) succinctly states 7 assumptions and illustrates how to test them using R. These assumptions are:

Assumption 1: The observations are independent of each other.

Assumption 2: The response or dependent variable is interval-scaled or ratio-scaled.

Assumption 3: The relation between response and explanatory variables is linear.

Assumption 4: The variance of error terms is constant, i.e. the homoscedasticity of variance.

Assumption 5: No strong linear dependence between explanatory variables (multicollinearity).

Assumption 6: The error terms are not autocorrelated.

Assumption 7: The error terms are normally distributed, with a zero mean.

Detailed approaches to test whether any of these assumptions is empirically violated will be illustrated in sub-section Verifying the Regression Assumptions with R.

Spreadsheeting plus Statistical Computing with R

Associating quantitative linguistics with corpus linguistics in a more specialized manner, Gries (2017) step-by-step explains the state-of-the-art techniques of R for practical quantitative analysis in corpus and empirical linguistics. R is adopted in the present research because it is *ubiquitously used by numerous quantitative analysis research, including linguistics*. Here we need to emphasize that such R programs are, in fact, very small relative to programs written in other languages like C++ and Java. The integrated development environment (IDE) of R—*RStudio* is used in the present research. Even further, Unwin (2015) concentrates on *visualization*, one of the greatest strengths of R, and introduces a package of *3-D scatterplot*. Apart from R, spreadsheeting can indeed be capitalized on. And our research experiences fruitful complementary use of them.

Among linguists, Gries is a prolific quantitative linguist who articulates a multitude of books and journal articles on this field. Gries (2015b) elucidates the use of R for statistical computing in empirical modeling in quantitative linguistics. Gries (2013) specializes in statistical tests for the analytics of textual learner corpus data. More broadly, Gries (2015a) comprehensively consolidates key *statistical methods* for learner corpus research (LCR). Specifically, Gries & Deshors (2015c) apply *regression modeling* and urge that rigorous methodology like this should be used.

Baclawski (2008) uses R to demythologize a wide range of topics in probability, including "sets, events and probability"; "finite processes"; "discrete random variables"; "general random variables"; "statistics and the normal distribution"; "conditional probability"; "entropy and information"; and "Markov chain" (a stochastic process which is "based on sequences of dependent random variables but for which the dependence is of the simplest possible kind: *the future depends on the present but not on the past*" pp. 303) etc. And all these topics are illustrated by using R, which is a very popular language widely adopted in quantitative linguistics.

More generally, linear regression, no matter simple (only one explanatory variable) or multiple (at least two such variables), is a kind of *linear statistical models*, computing of which is explicated by Faraway (2015) using R. The *trend of proliferating statistical computing* adopted in many contemporary texts show that R is the statistical computing 'package' most widely used throughout some texts although such implementation is not explicitly expressed in books' titles.

Among statistical computing texts, Crawley (2015) and Fox & Weisberg (2011) offer practical application of R to *basic statistics* and to *regression* respectively. For up-to-date and encyclopedic consultation, readers are referred to Ugarte, Militino & Arnholt (2016), who offer a balanced and lucid exposition of probability and statistics in terms of *mathematical notations* and of *R modeling*.

Computational Approach to Concgramming

A still young innovation in MWE research developed in the last decade is the concept concgram, which is the linguistic basis for the implementation of ConcGram©—a phraseological search engine. The concgram is a *linguistic conception*, while ConcGram© is the software discovering instances of it from corpora automatically. Linguists and researchers of *The Hong Kong Polytechnic University* collaboratively conduct researches into concgramming, which include many insights from giant corpus linguist and lexicographer Sinclair. The present research adopts ConcGram© to extract all the instances of 2-word, 3-word, 4-word, and 5-word concgram-based MWEs, of which the counts of such instances of word tokens are their raw frequencies.

Phraseological Search Engine ConcGram©

One point that is particularly worthy of further consideration is the counting of raw frequency of each concgram, not the individual words within it, is currently done both fully automatically by using ConcGram© and semi-automatically by using Microsoft Word. In other words, such *raw counts are not absolutely precise but the precision is sufficient to demonstrate the substantiality of the concgrams in the present research*. On the other hand, the actual frequency of *each word in a concgram* is *directly counted* by ConcGram© and therefore is an exact, not rough, figure.

To manifest the power of ConcGram© in extracting concgram-based MWEs, i.e. co-occurred contiguous, non-contiguous, or hybrid phraseological sequences, the present research uses it to search those 52 coherence phrases collected from website Smart Words (2013). These phrases are proved authentically present in ICLE. And the present research does not follow the convention adopted by Greaves (2009) to specify a concgram, e.g. "in/other/words", because although these three words can occur in *various combinations of orders*, each of the 52 MWEs used in the learners' argumentative writing follows a *linear order from left to right* of this sequence of occurrence of tokens. Such left-to-right extraction can be easily specified in ConcGram©.

Because the present research shares and agrees with the assertion that an MWE is under constant evolution in use and reality, we aim to, as automatically as possible, draw induction of such patterns from some authentic and naturally occurring texts produced by L2 users in the 6 sub-corpora of ICLE. In fact, astronomical amount of corpus data makes it *pragmatically infeasible for linguists to manually distil generalizable patterns* by making induction from the instances generated by ordinary concordancers.

In this study, MWEs are obtained from website Smart Words (2013), instead of being fully automatically unearthed by ConcGram©. The list is not authoritative but this does not matter because what we need is *a list of MWEs as stimuli* that are to be *checked against the authentic contents of ICLE by ConcGram©*.

ConcGram© is offered by Greaves (2009). What concgramming is can be succinctly specified: it is an *analytical process that helps the user to identify variations within phraseological sequences called concgrams*. ConcGram© is the choice of the present study to conduct *corpus-based* empirical analytics in the sense that concgrams are *pre-specified* as per website Smart Words (2013), though it can carry out corpus-driven empirical analytics. This software can automatically and dynamically extract concgram-based multi-word expressions (MWEs) no matter they are contiguous (e.g. "for instance"), non-contiguous (e.g. "under…circumstances"), or hybrid (e.g. "not only…but also").

Corpus-based vs Corpus-driven Searching

But our preliminary use of ConcGram© also finds that if users do *not*, in advance, *pre-specify* the patterns of MWEs to be discovered, the concgramming process can be very longish because there are *tremendous permutations of word orders* to be trialed by ConcGram©. And sometimes some meaningless sequences of words are unearthed by the software, while sometimes users need to *eyeball the outputs to differentiate genuinely meaningful phraseological sequences from the mistaken*. That means the *benefits of a completely flexible identification can be offset by the costs of manual inspection* of the mistaken.

Without pre-specifying concgram-based MWEs in the 6 sub-corpora in ICLE for searching, users would experience that searching can be highly *time-consuming* according to our preliminary trial of using ConcGram©, which may continuously run for *many hours or even 1 day* but *not meaningful 'collocations' may be discovered*. On the contrary, checking the list of coherence phrases offered on website Smart Words (2013) against the corpus data is *highly efficient* and helps confirm that these phrases are also ubiquitously used by learners of Standard English. Thus, the present research *pre-specifies* the MWEs for ConcGram© to search them in ICLE.

Strength

Phraseological search engine ConcGram© version 1.0 is an amazing software package. It is already a very versatile tool for *autonomous or pre-specified extraction* of contiguous, non-contiguous, and hybrid co-occurrence of words, i.e. lexical items or meaning shift units (MSU) in corpus linguist and lexicographer Sinclair's words. Along the lines of concgramming, Cheng & Leung (2012) highlight that *non-contiguous phraseology* is not as widely investigated as the n-gram and the skipgram (younger than the n-gram) in classical framework of phraseology. They realize various patterns of *phraseological (constituency and positional) variations* by exemplifying concgrams extracted from corpora with the help of ConcGram©. This software can distill all "the *specific patterns* of all possible phraseological variations"—concgrams, even though the *frequencies* of concgrams may need to be counted semi-automatically with the aid of other software.

In spite of its flexibility in searching can create burdens on users, who need to eyeball the outputs to judge whether they are truly MWEs, ConcGram© *transcends the classical concept—the n-gram*, which is *only a contiguous sequence of words*. Indeed, it automatically identifies the variations of a sequence of words in a concgram, e.g. "the expenditure of the government" versus "government expenditure". Therefore, the present research does not negate the contribution of ConcGram©, but because it is not open-source, no further enhancement can be done to this software by someone, except its creator.

Limitation

In principle, concgramming is based on raw frequencies gauged *computationally* rather than *inferential statistics*. Sinclair points out that "the practice of calculating the significance of the collocation" of two words "by comparing its frequency with the respective frequencies of" these two words "in the…corpus as a whole is *wrong*". This is a feature but may be a characteristic deficiency. On the one hand, it is this computational methodology that ConcGram© distinguishes itself; on the other hand, it is this methodology that isolates ConcGram© from statistical counterparts. Although bivariate mutual information can be computed by ConcGram©, ConcGram© emphasizes the *computational paradigm while neglecting the potential benefits and power of more sophisticated statistical techniques, especially multivariate mutual information* (MMI).

Research Questions and Hypotheses

Under the overarching bottom-up approach, the research questions and hypotheses of the present research are not set out until now. These two research questions will be empirically ascertained in section Empirical Analytics and Discussion. *For brevity, variable name MWE Length rather than concgram-based MWE Length is hereafter used.*

1. Are Shannon Entropy and MWE Length empirically positively related?
   H_0: Shannon Entropy and MWE Length are <u>not</u> statistically positively related.
   H_1: Shannon Entropy and MWE Length are statistically positively related.

2. Are Shannon Entropy and Markov multivariate mutual information (MMMI) empirically inversely related?
   H_0: Shannon entropy and MMMI are <u>not</u> statistically inversely related.
   H_1: Shannon entropy and MMMI are statistically inversely related.

Empirical Analytics and Discussion

This section first depicts the evidence of linguistic phenomenon—multi-word expressions (MWEs). And then it presents concrete instances of a specific concgram to back up the claim that the evidence is authentic. Next, it summarizes the numerical key indicators/variables of the present research, i.e. MWE Length, Markov MMI, and Shannon Entropy. Further, it explicates what hypothesis testing, p-values, and effect sizes are. Moreover, it demonstrates a distinctive feature of R: scatterplot matrix of variables in regression. Besides, it constructs an empirical multiple linear regression model. On top of that, it provides the experimentation with R and interpretation of outputs. Afterward, it illustrates how to identify outliers and overmuch influential observations with R and how to verify the regression assumptions with R. Finally, it scrutinizes the empirical findings on the two research questions and hypotheses.

Evidence of Linguistic Phenomenon: MWEs

Available on website Smart Words (2013) is a pool of formulaic phrases, i.e. a kind of multi-word expressions (MWEs) fixed in structure, that is selectively reproduced in the table below. Although that website is not an authoritative source, most of these MWEs are *empirically proved significantly present in the International Corpus of Learner English (ICLE)*, which the whole research is based on and driven by. The table also contains some concgrams that come from *intuitive usages*. But they and those in the pool are all screened such that each of them has non-zero frequency in any of the 6 ICLE sub-corpora with 100% argumentative writing. And MWEs are deliberately chosen such that they are *at least 2-word long but at most 5-word so as to fully exploit the strength of ConcGram*© to process MWEs. The present research believes that most of the educated users of Standard English consider these 52 formulaic MWEs idiomatic.

Concgram	**Website Evidence of Formulaic MWEs**
2-word Concgram (Contiguous)	“above all”, “after all”, “because of”, “claim(s) that”, “compared to”, “compared with”, “due to”, “even if”, “even though”, “for example”, “for instance”, “in addition”, “in conclusion”, “in fact”, “in general”, “in reality”, “in short”, “of course”, “suggest(s) that”, “that is”, “this means…”, “to conclude”, “to explain”.
2-word Concgram (Non-contiguous)	“both…and…”, “either…or”, “for...reason”, “rather…than”, “should not”, “so…that”, “to…extent”, “under…circumstances”.

3-word Concgram (Contiguous)	"as a result", "as long as", "as soon as", "as well as", "in my opinion", "in order to", "in other words", "in that case", "on the contrary", "so as to", "to sum up".
3-word Concgram (Hybrid)	"as much...as", "whether…or not".
4-word Concgram (Contiguous)	"at the same time", "in the case of", "in the first place", "in the same way", "on the other hand".
4-word Concgram (Hybrid)	"not only…but also", "as far as…concerned".
5-word Concgram (Contiguous)	"as a matter of fact".

Authenticity of Evidence

The list on the next page shows 67 authentic instances of 4-word concgram 'as far as…concerned' that are extracted by ConcGram© from the aggregate sub-corpus of the 6 ICLE sub-corpora of learner English written by learners of 6 L1s—Bulgarian, Chinese, Japanese, Russian, Tswana, and Turkish. ConcGram© is instructed to *extract this pattern rightward only rather than both leftward and rightward*. That means, for instance, the phrase 'concerned' could appear on the left as well as on the right of 'as far as', but now ConcGram© does not extract those instances with 'concerned' on the left-hand side. The raw frequency of instances (i.e. tokens) of each single word type in the MWE above, i.e. 'as', 'far', 'as', and 'concerned', is counted and generated by ConcGram© in a list of *unique words* (i.e. types).

As to the concgrams consolidated in the table in sub-section Evidence of Linguistic Phenomenon: MWEs, they are *empirically proved not trivial* in ICLE in terms of their Markov multivariate mutual information (MMMI). Again, let's take the 4-word hybrid concgram "as far as…concerned" as an example. The MMMI of this concgram in the aggregate sub-corpus (consolidated from the sub-corpora of the 6 L1s) is 6.88. This figure indicates that the association among the words in this concgram is *significant* although *no specific benchmark level is set* in the present research.

A point worthy of emphasis is that although 'as far as' and 'concerned' are structurally fixed, the whole MWE is dynamic since the *number of in-between words is flexible* as denoted by the symbol '…' and completely depends on the actual co-occurrences (in other word, corpus-driven) in each of the 6 sub-corpora mentioned above and in their aggregate. For instance, the 67 instances of this MWE "as far as…concerned" are listed below.

1 of their privileges and have to work as much as men, and as far as the low class is concerned,
2 the roads may get damaged all the way through. As far as I'm concerned, I may say the following :
3 form of imagination - the material realization. As far as I'm concerned this is definitely so.
4 troubles the baby might be faced with if born. As far as i'm concerned, to preserve the baby from
5 we have some unpleasant associations with it. As far as I'm concerned, they remind me of the myth
6 below you feel they don?t have to answer to you. As far as I?m concerned people who still think that
7 there must be various reason about that. But as far as I'm concerned, I stand in opposition to a
8 on reading rather than on listening or speaking. As far as I am concerned, substantially, there was
9 that of Development studies, Peace studies, etc. As far as I??m concerned these degrees should be
10 cannot say that money is the root of all evil. As far as I am concerned I cannot help admiring such
11 do. It is the waste of their time and life. As far as I am concerned, people should understand
12 may poison the teenagers' mind. However, as far as I am concerned, I do think television does
13 to live like they live, or to feel what they do. As far as I am concerned I always try to put myself
14 spread over the disease. As far as I am concerned, poverty is not the cause
15 economic would be suffering a serious problem. As far as I am concerned, banning smoking in a
16 the undergraduates were much worse than before. As far as I was concerned, I do not agree with them.
17 a TV now. A TV can accompany us for many years. As far as this is concerned, it is quite practical
18 country but European country. Why Our leaders as Far as Fifa its concerned why dont they ask them
19 role which TV plays in our every-day life. As far as cinema is concerned, the picture is
20 cynical producers the answer can be affirmative. As far as drugs are concerned, we have a strict
21 children's behaviour and achievements. As far as pupils are concerned, the many they get at
22 - there is a blind masseur in our health centre. As far as charity is concerned it is an altruistic
23 less sport, less care for ourselves and others. As far as dreaming is concerned, perhaps the most
24 all these things somewhere, but not in the army. As far as the army is concerned it must consist
25 So men generally were there main culprits as far as feminism is concerned. Even in places like
26 that these people have dreams which guide them. As far as industry is concerned, it is not so recent
27 any big city like Moscow, London or New York. As far as New York is concerned, I think , it is a
28 who are not satisfied about the way our leaders as far as soccer it is concerned Many people are not
29 of casualities (or without then at all). So, as far as my opinion is concerned, I suppose that
30 from reality or have great influence on us.As far as believers are concerned I do not think
31 opinion of this controversy. In the first place, as far as many parents' concerned, the television is
32 in books. Anyone has this kind of imagination. As far as the latter is concerned, it is in charge
33 them to survive in such a difficult world. But as far as television is concerned, I'm quite
34 as good and evil, love, kindness and humanity. As far as imagination is concerned poetic literature
35 The pollution of owls and snakes fell and so on. As far as Agriculture is concerned it is another big
36 if it've already committed then how to react. As far as first point is concerned, the attention of
37 for instance, in France and also in Russia. But, as far as this system is concerned, we have to, I
38 he be able to reap the fruit of his labour ? As far as this aspect is concerned , I am inclined
39 to avoid the tragedy, avoid unnecessary victims. As far as the Gulf War is concerned it has shown us
40 view of "the world" appears to be more important as far as our existence is concerned. And now let us
41 people come on this world equal of each other as far as human rights are concerned. Every living
42 can be regarded as a major principle of harmony as far as this relation is concerned. What is more,
43 privileges and have to work as much as men, and as far as the low class is concerned, even on
44 the number of thefts can be reduced. But as far as graver crimes are concerned, such as
45 is concerned, my expectations are fulfulled. But as far as the curriculum is concerned there is much
46 valuable education, but I am rather optimistic as far as this situation is concerned. I think that
47 they were practical and generally agreed upon. As far as modern feminism is concerned, it seems to
48 prisons... Who cares, what's wrong there now? As far as western society is concerned, the
49 performances, meetings with the famous people: As far as elderly people are concerned they, in
50 Noone can argue with this. But this is true only as far as material values are concerned: everyone
51 sometimes do more harm than good for humanity. As far as artistic people are concerned it is their
52 so quickly as in this century. And never, as far as our civilisation is concerned, a man came
53 shelter and clothes. Food is very important as far as physical fitnees is concerned. Player
54 that God does exist. It is their own business. As far as religious rites are concerned now they do
55 took liberties to claim the female's capacity as far as the power of mind was concerned. I would
56 It is exactly the same thing, or even worse, as far as capital punishment is concerned. Fifteen
57 Institutions in Africa, there's still ignorance as far as hiv aids education is concerned. People
58 them all very different and very interesting. As far as younger generation is concerned, I can
59 has a lot to do with dreams and imagination, as far as the great scientists are concerned. But
60 jobs and don't want to lose them. And as far as really talented women are concerned (that
61 university I expected it to be the same with me. As far as life outside university is concerned, my
62 your hair is blue or yellow, long or short. As far as fashion of modern people is concerned
63 lag behind the children of some other countries as far as their mental development is concerned.
64 of the country, are of a much greater value as far as their creative potential is concerned,
65 their students, which is rather a disadvantage as far as learning foreign languages is concerned.
66 contribution and money are incommensurate. And as far as our daily work for society is concerned,
67 suffer from the consequences of feminism. And as far as my subjective point of view is concerned,

Numerical Summary of Key Indicators

Here is a summary of all the numerical values of the key indicators—MWE length, MMMI and Shannon entropy—of the *aggregate sub-corpus derived from the 6 sub-corpora with 100% argumentative writing* in ICLE. These indicators are to be addressed in the multiple linear regression modeling. For the mechanisms to obtain MMMI and Shannon entropy, readers are referred to sub-section Markov Multivariate Mutual Information (MMMI) and Shannon Entropy as

an Uncertainty Measure. Let this report not get bogged down with too much data, only the data of the aggregate sub-corpus but not that of each of the 6 sub-corpora is provided here.

Concgrams	MWE Length	Markov MMI	Shannon Entropy
“above all”	2	3.86	0.0402
“after all”	2	3.86	0.0496
“because of”	2	2.44	0.2245
“both…and…”	2	4.29	0.1807
“claim(s) that”	2	5.08	0.1137
“compared to”	2	3.51	0.1993
“compared with”	2	6.03	0.0499
“due to”	2	4.45	0.2023
“either…or”	2	6.61	0.0508
“even if”	2	4.41	0.0605
“even though”	2	6.99	0.0253
“for example”	2	5.94	0.0994
“for instance”	2	4.96	0.0930
“for...reason”	2	3.08	0.0941
“in addition”	2	5.32	0.1500
“in conclusion”	2	4.28	0.1504
“in fact”	2	4.18	0.1554
“in general”	2	3.78	0.1485
“in reality”	2	2.18	0.1499
“in short”	2	3.75	0.1487
“of course”	2	4.55	0.1953
“rather…than”	2	7.84	0.0251
“should not”	2	3.33	0.1106
“so…that”	2	2.81	0.1496
“suggest(s) that”	2	4.97	0.1143
“that is”	2	0.85	0.2563
“this means…”	2	3.36	0.0672
“to conclude”	2	4.09	0.1999
“to explain”	2	3.53	0.1992
“to…extent”	2	4.68	0.1995
“under…circumstances”	2	8.69	0.0064
“as a result”	3	6.85	0.2058
“as long as”	3	6.95	0.0777

Concgrams	MWE Length	Markov MMI	Shannon Entropy
“as much...as”	3	6.34	0.0845
“as soon as”	3	6.99	0.0712
“as well as”	3	7.17	0.0797
“in my opinion”	3	5.05	0.1736
“in order to”	3	5.40	0.1630
“in other words”	3	5.31	0.1828
“in that case”	3	5.50	0.2653
“on the contrary”	3	6.00	0.3364
“so as to”	3	6.73	0.1076
“to sum up”	3	4.96	0.2077
“whether…or not”	3	9.72	0.0635
“as far as…concerned”	4	6.88	0.1291
“at the same time”	4	7.76	0.3507
“in the case of”	4	4.51	0.6170
“in the first place”	4	4.93	0.4484
“in the same way”	4	4.94	0.4550
“not only…but also”	4	6.65	0.1812
“on the other hand”	4	7.14	0.3646
“as a matter of fact”	5	7.06	0.4002

Hypothesis Testing, P-values and Effect Sizes

As for hypothesis testing, Ugarte, Militino & Arnholt (2016 pp. 522) elucidate it in the table below:

		Statistical Decision	
		Reject H_0	Fail to Reject H_0
Null Hypothesis	True	Type I Error P(Type I Error) = P(reject H_0 | H_0 is true) = α, which is the level of significance	Correct Decision P(Accept H_0 | H_0) = $1 - \alpha$, which is the confidence level
	False	Correct Decision P(Accept H_1 | H_1) = $1 - \beta$, which is the power of the test	Type II Error P(Type II Error) = P(Fail to reject H_0 | H_0 is false) = β

Crawley (2015 pp. 3-4) introduces that *significance means "a result was unlikely to have occurred by chance"*. An agreed convention among statisticians is that "an event is unlikely if it occurs less than 5% of the time". Great philosopher Karl Popper once argued that "a good hypothesis is a *falsifiable* hypothesis". One of the big notions in the philosophy of science is that *"absence of evidence is not evidence of absence"*. Take regression modeling as an example. The null hypothesis is that "y is not a function of x, or y is independent of x". The key is that "*the null hypothesis is falsifiable*" and can be rejected when the data reveals that the null hypothesis is "sufficiently unlikely".

A "much-misunderstood" notion is *p-value*, which is not the probability that the null hypothesis H_0 is true. But it is "calculated on the assumption that the null hypothesis is true". Thus, H_0 can be *rejected* when the corresponding p-value is very small, often less than 0.05. "*p-values are about the size of the test statistic*", e.g. "Student's *t*, Fisher's *F* and Pearson's chi-squared". "*Big* values of the test statistic indicate that the null hypothesis is *unlikely to be true*." Lastly, *p*-values are indicative, but "effect sizes and sample sizes are equally important in drawing conclusions. The modern practice is to *state the p-value rather than just to claim: we reject the null hypothesis*."

According to Levshina (2015 pp. 103), "statistical significance does not tell you anything about the *effect size*." Levshina (ibid. pp. 129-130) expounds that "an effect size shows *how strongly* different variables are related/associated, or *how greatly* groups of observations differ from one another. The correlation coefficient r is a good example of effect size. *Statistical significance*, which is associated with the *p*-value, does not show the *strength* of a relationship or the *magnitude* of a difference. It *only shows how confident one can be that the observed relationship or difference are not due to chance alone*. *A strong effect does not automatically entail significance, and vice versa*. Crucially, if the same effect size is observed in a smaller sample and a larger sample, the *p*-value will be smaller in the latter." "If there is a correlation in the population, the chances of detecting it in the data increase with the sample size."

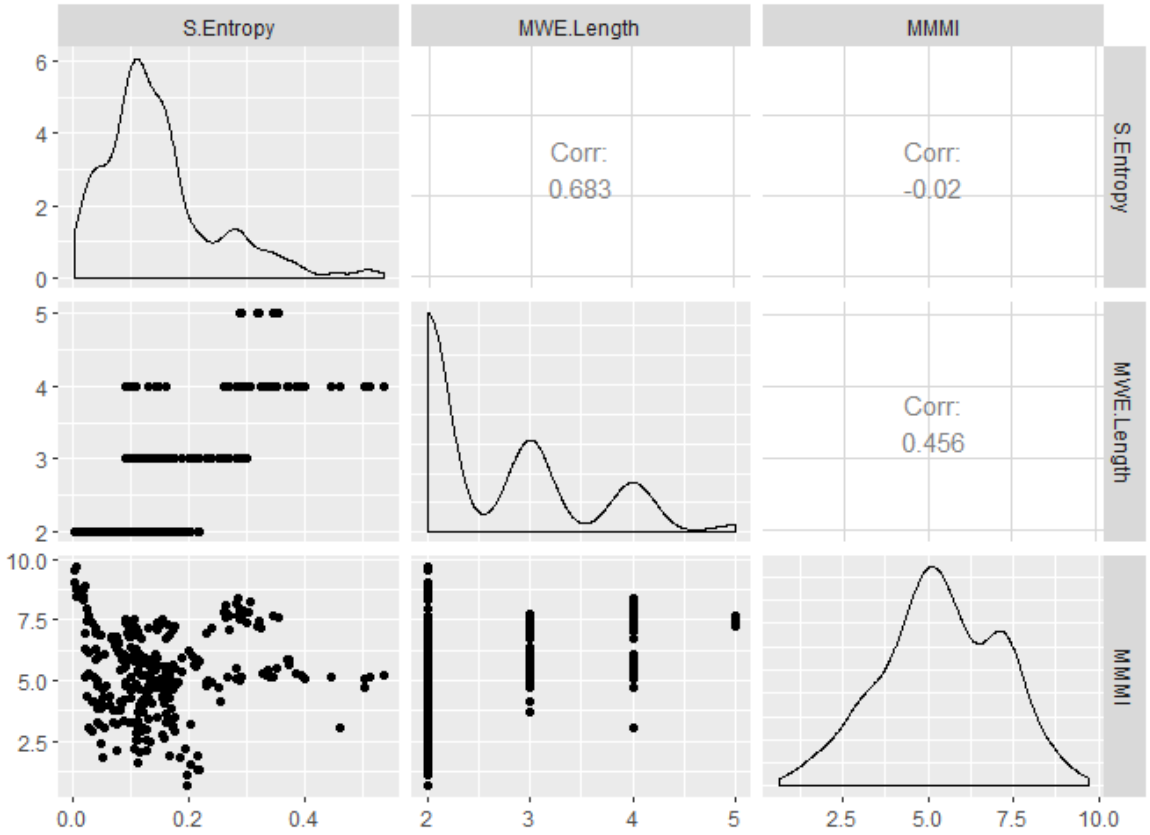
Scatterplot Matrix of Variables in Regression with R

To facilitate *reproducible research* by other linguists and researchers, R codes for loading the necessary libraries and the source data file as well as for producing the scatterplot matrix of variables in regression are provided below:

```
library("gdata", lib.loc="~/R/win-library/3.4")
library("car", lib.loc="~/R/win-library/3.4")
library("scatterplot3d", lib.loc="~/R/win-library/3.4")
library("Rling", lib.loc="~/R/win-library/3.4")
```

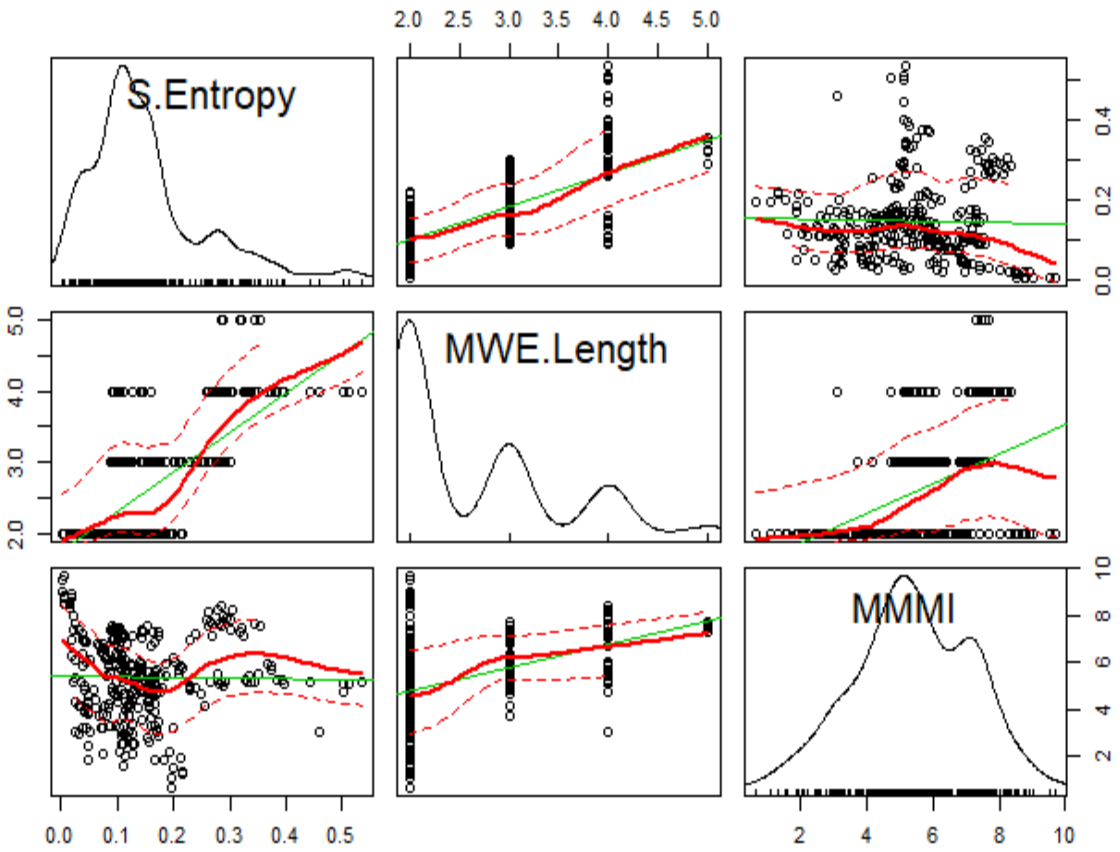
```
library("GGally", lib.loc="~/R/win-library/3.4")
ds1 = read.xls("f:\\LT6580 Master Project\\ICLE Empirics\\Argumentative
      100\\argumentative100.xlsx", sheet = "Key Variables")
attach(ds1) # ds1 is a name of the data set defined a user
ggpairs(ds1, diag = list(continuous="densityDiag", discrete="barDiag"),
        axisLabels = 'show')
scatterplotMatrix(~ S.Entropy + MWE.Length + MMMI)
```

Those variables in the empirical multiple linear regression to be explicated in the next section are S.Entropy (i.e. Shannon Entropy), MWE.Length (i.e. MWE Length), and MMMI (i.e. Markov Multivariate Mutual Information). Let us illustrate how to interpret this scatterplot matrix. How the data series of Shannon Entropy is correlated with MWE Length and MMMI is plotted from left to right in the scatterplot matrix shown below. On the topmost row, the leftmost figure shows the *distribution of Shannon Entropy*, the next two on the right shows how Shannon Entropy is *correlated* with MWE Length and with MMMI (i.e. 0.683 and -0.02). On the bottom row, the rightmost figure shows that MMMI is *approximately normally distributed* (i.e. bell-shaped). Since normal distribution is a *continuous* probability distribution, it cannot be applicable to MWE Length, which is a *discrete* random variable.



On another scatterplot matrix shown below, the *first row* of diagrams is our focus. There is a salient *positive relationship* between the response variable Shannon Entropy and the explanatory

variable MWE Length, as the upward sloping curve reveals. Besides, there is a *negative relationship* between Shannon Entropy and MMMI, as the downward sloping curve indicates.



According to Levshina (2015 pp. 16-18), independent variable MWE Length is *quantitative, discrete and interval* (i.e. 2 words < 3 words < 4 words < 5 words, in which the interval is constant and comparable), and another independent variable MMMI is *quantitative, continuous and ratio* (which is the same as "interval variable" but includes zero within the interval). And dependent variable Shannon Entropy is also *quantitative, continuous and ratio*.

Consolidating the analyses above, these two scatterplot matrices offer comprehensive views of the relationships between the dependent variable Shannon Entropy and the explanatory variables, MWE Length and MMMI. Such *pre-regression quantitative analyses* are crucial.

Empirical Multiple Linear Regression

In practice, one of the *empirical prerequisites* for valid application of *inferential statistical models* is that the number of points in data series should not only be larger than a minimum threshold (a practical rule-of-thumb) but also be as large as data is available. Therefore, our research examines

all the 6 sub-corpora of 100% argumentative writing in ICLE. These 6 sub-corpora together have 1.12 million words. But this does not mean that there are 1.12 million data points; *data points* are those observations of 2-word concgrams (e.g. "because of"), 3-word concgrams (e.g. "as a result"), 4-word concgrams (e.g. "in the same way"), and 5-word concgrams (e.g. "as a matter of fact"). The *sample size is of 312 observations*.

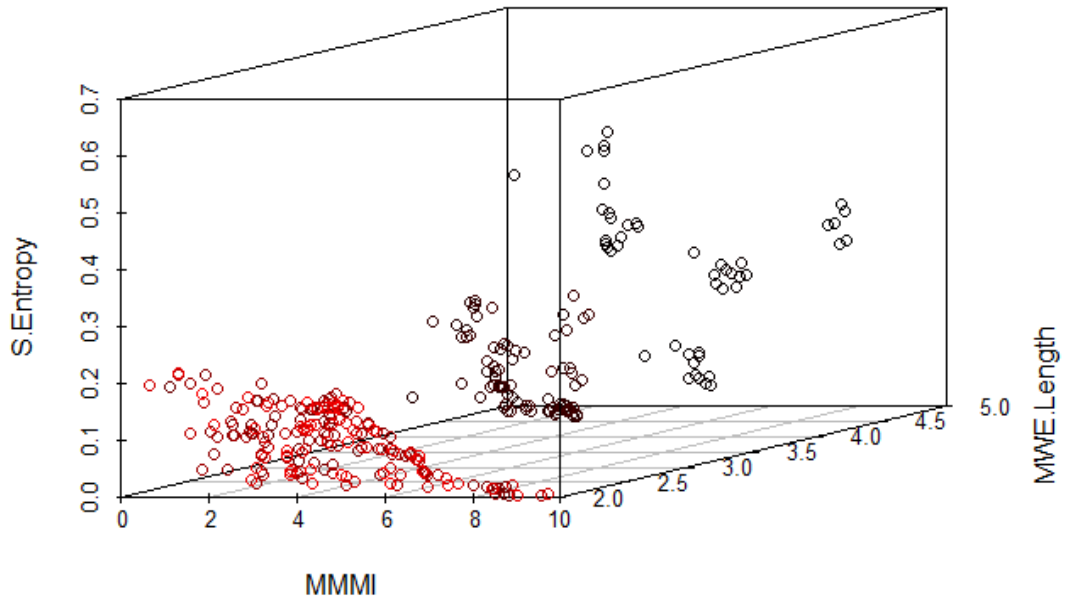
Among the concgram-based MWEs on website Smart Words (2013), 5-word concgrams most concern us because the existing findings are based on only one significant 5-word concgram as other 5-word concgrams on this website, e.g. 'to put it another way', are finally dropped out of our list as they do *not occur in all* the 6 sub-corpora with 100% argumentative writing in ICLE.

After extensive experimentation with the authentic data on ICLE extracted by ConcGram©, a parsimonious empirical multiple linear regression model emerges:

$$S.Entropy_i = \beta_0 + \beta_1 MWE.Length_i + \beta_2 MMMI_i + \varepsilon_i$$

In theory, the relations among Shannon Entropy, MWE Length and MMMI are deterministic, but *empirically* such relations are *stochastic* as there are *random noises* ε_i interfering the pure linear relation among these three variables. Moreover, Crawley (2015 pp. 200) suggests that the above analyses enable linguists to start multiple linear regression modeling.

Experimentation with R and Interpretation of Outputs



The *3-D scatterplot* above shows that Shannon Entropy and MMMI is *inversely related* and is consistent with our expected result—the construct entropy represents uncertainty and mutual information represents the *reduction in uncertainty* about one random variable, say a word, given the information/knowledge of the other, say another word. The higher the MMMI, the lower the Shannon Entropy, and vice versa. That is, *increased information implies decreased uncertainty*.

On the other hand, Shannon Entropy and MWE Length is positively related. That is, as MWE Length increases, Shannon Entropy also increases. Why? It is mainly due to the formulaic specification of Shannon Entropy (recapped below):

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Since $\log_2 p_i$ must be negative but additional variables' probabilities are added to the cumulative total, which is made positive by the negative sign. That means, H will increase accordingly. Thus, that Shannon Entropy increases as MWE Length increases can be *logically inferred*.

To facilitate reproducible research by other researchers, the R codes for generating the 3-D scatterplot on the previous page are shown below:

```
par(mfrow = c(1,1)) # 1 row, 1 column
scatterplot3d(MMMI, MWE.Length, S.Entropy, highlight.3d = TRUE, angle = 24)
```

The following R codes also facilitate reproducible research: `lm` is a R function that conducts linear regression modeling, and `summary` is a R function that generates inferential statistics on the model. 'mBasics' is the user-defined name of the model used in the present research.

```
mBasics <- lm(S.Entropy ~ MWE.Length + MMMI, data = ds1)
summary(mBasics)
```

As per Levshina (2015 pp. 145), the summary generated by R below first "repeats the formula" of the basic empirical model specified above. Next item is *Residuals*, which "should be normally distributed and center around zero." Levshina (2015 pp. 162) recommends using the "Shapiro-Wilk" normality test to check" whether the assumption normality is violated. This test is to be conducted in sub-section Verifying the Regression Assumptions with R.

Then, another set of statistics is a table of *Coefficients*. Levshina (ibid. pp. 145) states that "the first column shows the estimates, which specify the intercept and the slopes of the regression line." "In addition, the table displays the standard errors of estimated coefficients." "The *p*-values in the rightmost column are based on the *t*-statistics (see the second column from the right) and show how

confident we can be that the values are different from zero. A *p*-value less than 0.05 suggests that the null hypothesis of no effect can be rejected."

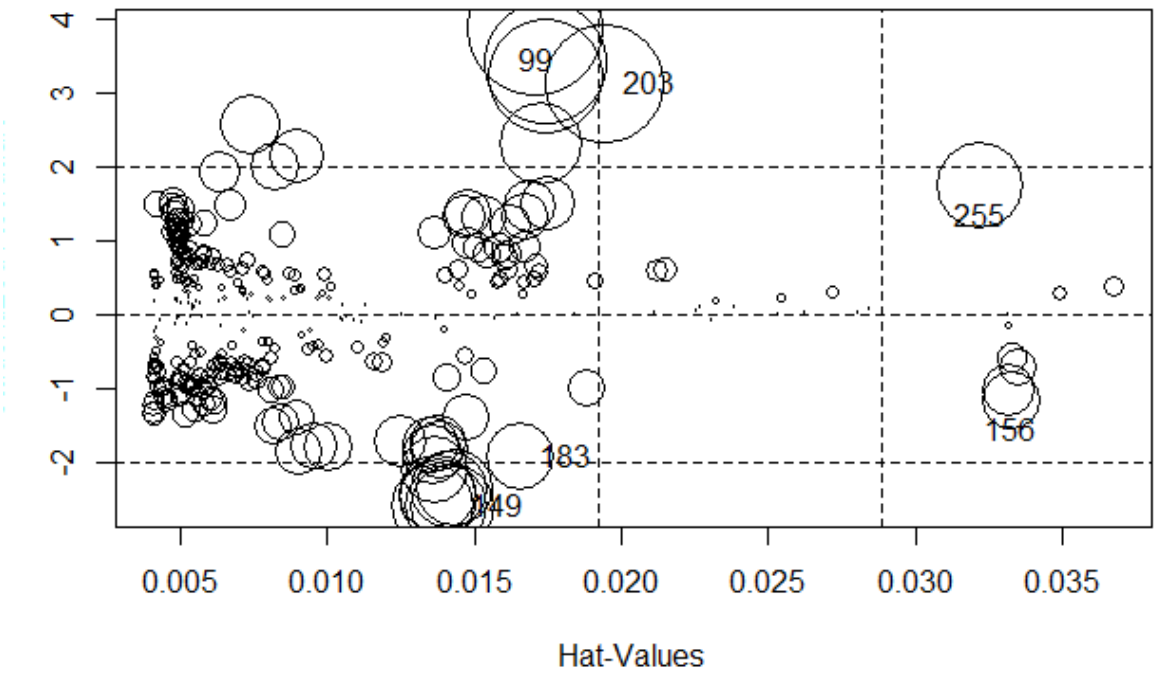
```
Call:
lm(formula = S.Entropy ~ MWE.Length + MMMI, data = ds1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.158425 -0.044415  0.006128  0.037259  0.230544

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.004575   0.013441   -0.34    0.734
MWE.Length   0.107519   0.004941   21.76   <2e-16 ***
MMMI        -0.023458   0.002250  -10.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06158 on 309 degrees of freedom
Multiple R-squared:  0.6053,   Adjusted R-squared:  0.6028
F-statistic: 236.9 on 2 and 309 DF,  p-value: < 2.2e-16
```

There are two significant explanatory variables MWE Length and MMMI since their *p*-values, under the heading `Pr(>|t|)`, are less than 0.05; and their corresponding null hypotheses stated in sub-section Research Questions and Hypotheses can be rejected. This is also indicated by *** right after their *p*-values. And R^2 is 0.6053, which means the 'model fitting' (i.e. the degree of fit) of this multiple linear regression is quite high. As per Crawley (2015 pp. 133), R^2 can be interpreted as "the *fraction* of the *total variation in y that is explained by the regression*." As Levshina (2015 pp. 149) suggests, `F-statistic` "shows if the model is significant in general. In most cases, this means that at least one variable has an estimate that is significantly different from zero. This statistic is the *ratio* of the *variance explained by the model and the residual variance*." This statistic "should be at least greater than 1, and in the present model, it is 236.9, which is very high. And "the *p*-value shows the probability of observing a given *F*-ratio if the null hypothesis were true (i.e., the variables had no effect)". Because of a very low *p*-value, it is concluded that the null hypothesis can be rejected and the model is significant.

Checking for Outliers and Overmuch Influential Observations with R

To account for the empirical findings of the regression above, discussion about the possible impacts of outliers and overmuch influential observations is necessary because as Levshina (2015 pp. 153-155) points out, outliers' "presence indicates a lack of fit" and overmuch influential observations "can—but not necessarily do—have a significant effect on the regression slopes, i.e. the estimates of regression *coefficients*". "To identify both types of problematic cases," the following R function can be used:

```
influencePlot(mBasics, id.method = "identify")
```



In the bubble plot above, the IDs are the *row numbers of the observations* among all the 312 observations in the data frame of the present research. This plot "is based on three values for each observation in the corpus: hat-values, studentized residuals and Cook's distances".

1. *Studentized residuals* (shown on the y-axis) are "normalized residuals", which are "adjusted by their expected variability." These residuals represent the discrepancies "between the actual and fitted values." It is recommended that in the bubble plot "*observations* with the values greater than 2 or smaller than -2 (on the *vertical axis*)" should be checked.
2. *Hat-values*, "the *x-axis*, indicate how much influence the observations can potentially have on the fitted values. The influence of a given observation is determined by the difference between the observation's value on an explanatory variable and the mean value of the variable. For leverage, there is no absolute threshold that needs to be controlled. The *vertical lines are drawn through the points which correspond to two and three times the average hat-values*."
3. *Cook's distances* "show the *effect of removing* an observation on the coefficients and fitted values. They are represented by the *size of the bubbles*."

Based on Studentized residuals, Hat-values, and Cook's distances, bubble 99, 149, 156, 183, 203 and 255 are identified as the outliers and overmuch influential observations in the data set of the present research driven by evidence from the International Corpus of Learner English (ICLE).

Verifying the Regression Assumptions with R

This section verifies the assumptions of using multiple linear regression models that are mainly based on Levshina (2015 pp. 153-162) and briefly based on Levshina (ibid. pp. 17).

Independence of Observations

This assumption "means that each value of the response variable should be independent from the values of other observations." As for the response variable Shannon Entropy as well as the explanatory variables MWE Length and MMMI, the values of their observations corresponding to a concgram-based MWE do not depend on those of other observations of another concgram-based MWE. Thus, this assumption is not violated.

Response Variable: interval- or ratio-scaled

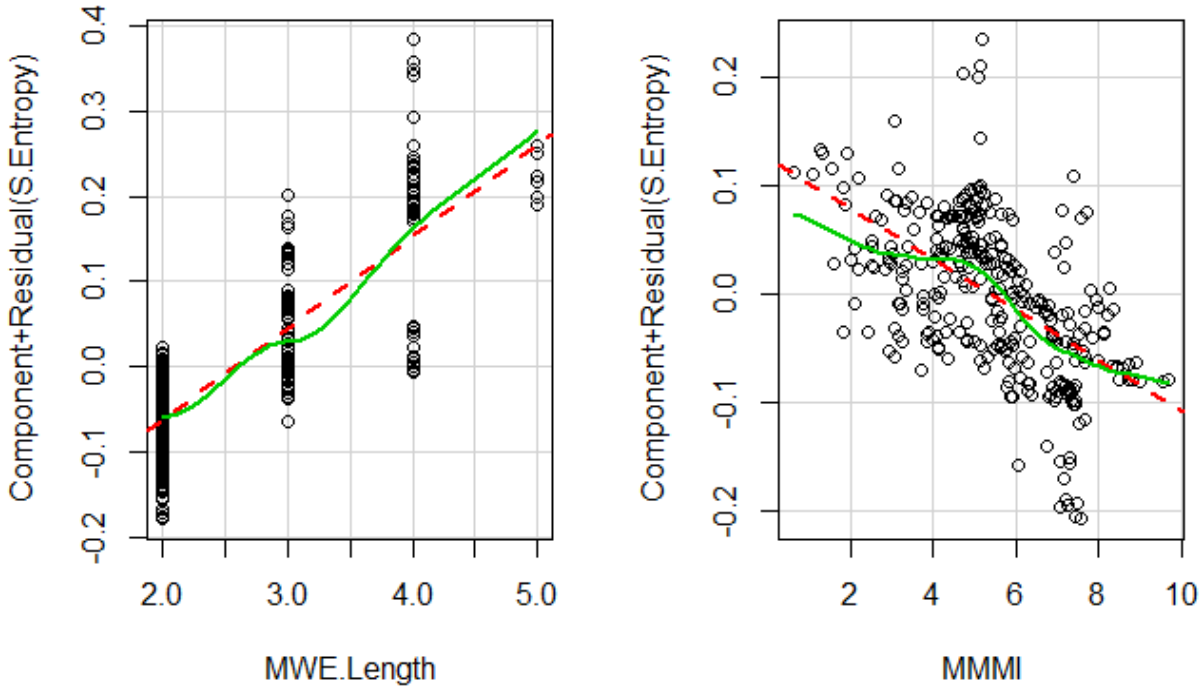
The response variable Shannon Entropy is clearly not only at least on an interval scale but in fact on a ratio scale, which is similar to an interval scale but also "includes zero on the scale, and the zero point is meaningful, not arbitrary."

Linearity between Dependent and Independent Quantitative Variable

"A convenient tool for detection of non-linearity in multiple regression analysis is the so-called component-residual, or partial-residual plot." The most 'basic' multiple regression model *without interaction terms and quadratic terms* is required by the R function `crPlot`, which does not accept such terms. The model specified in sub-section Empirical Multiple Linear Regression fulfills such a requirement. And MMMI is the only continuous, quantitative, explanatory variable in the model, while MWE Length is the only discrete, quantitative, explanatory variable.

```
par(mfrow = c(1,2)) # set the matrix of plots to be "1 row, 1 column"
crPlot(mBasics, var="MWE.Length")
crPlot(mBasics, var="MMMI")
```

"The component-residual plot shows how the residuals and the corresponding regression coefficient (the vertical axis) vary according to an explanatory variable (the horizontal axis) after "the effect of all other explanatory variables" is taken into account. In the present model, both MWE Length and MMMI are quantitative explanatory variables; the former is discrete while the latter is continuous. "The dashed red lines represent the slopes of the explanatory variables based on the regression estimates, and the solid green lines reflect the main tendency in the cloud of data points. The solid lines deviate from the dashed lines, indicating *some non-linearity*."

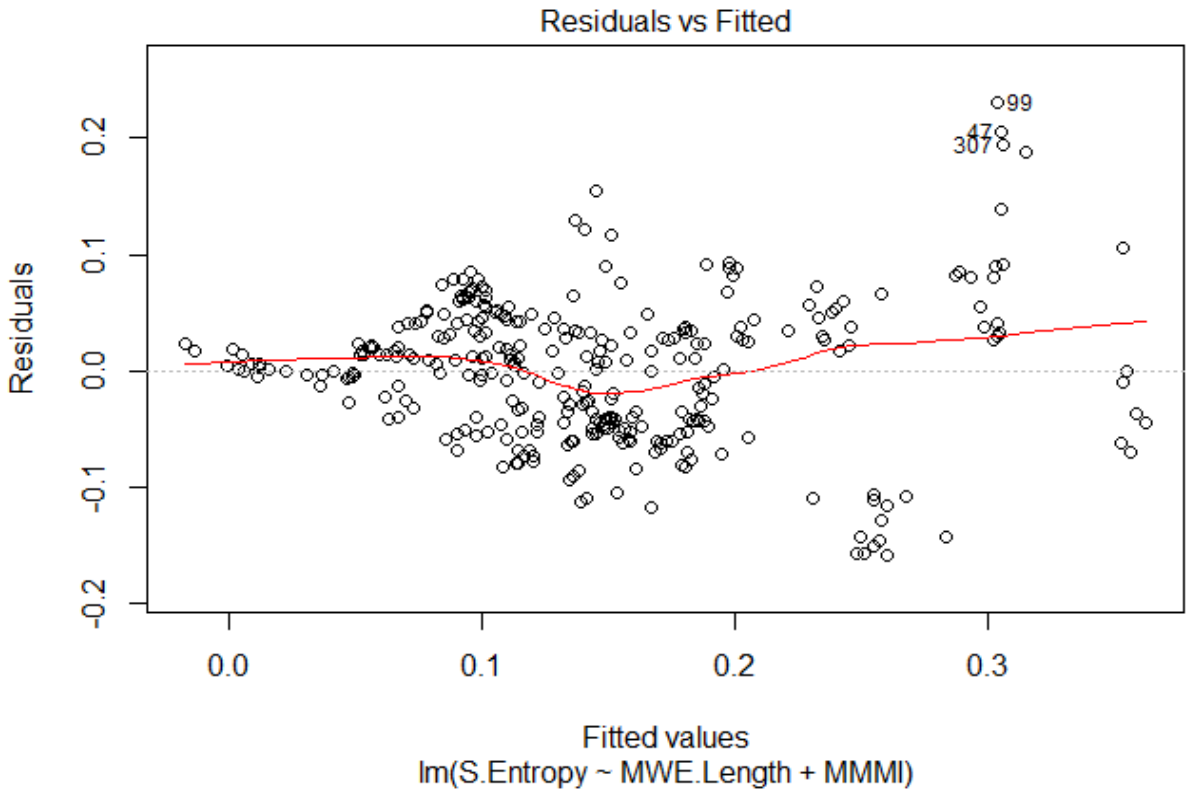


Homoscedasticity of Variance

"Visual diagnostic of homoscedasticity is possible with the help of the plot of the residuals against fitted values." Plotting a graph below is used to test out whether there is heteroscedasticity. This graph is generated by the following R codes:

```
par(mfrow = c(1,1)) # set the matrix of plots to be "1 row, 1 column"
plot(mBasics)
```

According to Crawley (2015 pp. 134-135), the graph "shows residuals on the *y* axis against fitted values on the *x* axis…It is a major problem if the scatter *increases* as the fitted values get bigger; this would show up like a wedge of cheese on its side (like this ◄…). That means there is heteroscedasticity." Seemingly, heteroscedasticity only *marginally exists* since there seems to be a ◄ shape but close inspection of the pattern seems to suggest that heteroscedasticity does *not exist* since the points on the rightward portion of the graph are sparingly dispersed. Since the pattern is *not very definite and clear-cut*, another test must be used to diagnose whether heteroscedasticity exists or not.



According to Levshina (2015 pp. 157), the *non-constant variance test* can also be used to check this assumption.

```
ncvTest(mBasics)
ncvTest(mBasics, ~MMMI)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 85.70169    Df = 1     p = 2.092342e-20

Non-constant Variance Score Test
Variance formula: ~ MMMI
Chisquare = 0.9143198    Df = 1     p = 0.3389708
```

"This test has the null hypothesis that the error has constant variance with the response (fitted values). The *p*-value smaller than 0.05 tells us that we can reject the null hypothesis of constant error variance." Since the *p*-value 2.092342e-20 of the *whole* model is significantly less than 0.05, the test rejects the null hypothesis and variance is non-constant. But a subsequent test for *individual* quantitative variable MMMI does not reject the null hypothesis. There is a dilemma: should the first model-wise *p*-value or the second variable-wise *p*-value be used as the decisive means to determine whether heteroscedasticity exists? From the point of view that MMMI is a quantitative explanatory variable, it is reasonably to make judgment based on the variable-wise *p*-value such that the whole

fitted model is marginally free from heteroscedasticity. But model-wise statistics should not be quietly neglected. Indeed, it is suggested that another kind of regression modeling called *semiparametric regression* is to be used for rectifying the above indeterminacy in future research. (see section Future Research for details)

No Multicollinearity among Explanatory Variables

"Multicollinearity is a phenomenon that can be observed when some variables relate to the same underlying causal effect." When explanatory variables that reflect one underlying factor are incorporated into a model, "very strongly correlated variables tend to have unstable estimates and large standard errors. That is, the estimates of correlated explanatory variables are no longer reliable, *even if the model has sufficient predictive power (i.e. a high R^2).*" The following R function returns *variance inflation factors*, "which help us estimate how much collinearity is associated with each regression term".

```
car::vif(mBasics)
```

```
MWE.Length       MMMI
  1.262641   1.262641
```

"There exist various rules of thumb, some of them stricter (VIF-scores should not exceed 5) and others less so (VIF-scores should not be greater than 10)." Since both VIF-scores are 1.262641, which is very small, multicollinearity does not exist in the model.

No Autocorrelation among Residuals

Autocorrelated residuals "tend to emerge when we have time series." The following R code can reveal whether autocorrelation among residuals exist:

```
durbinWatsonTest(mBasics)
```

```
lag Autocorrelation D-W Statistic p-value
  1       0.1254131      1.745709   0.018
Alternative hypothesis: rho != 0
```

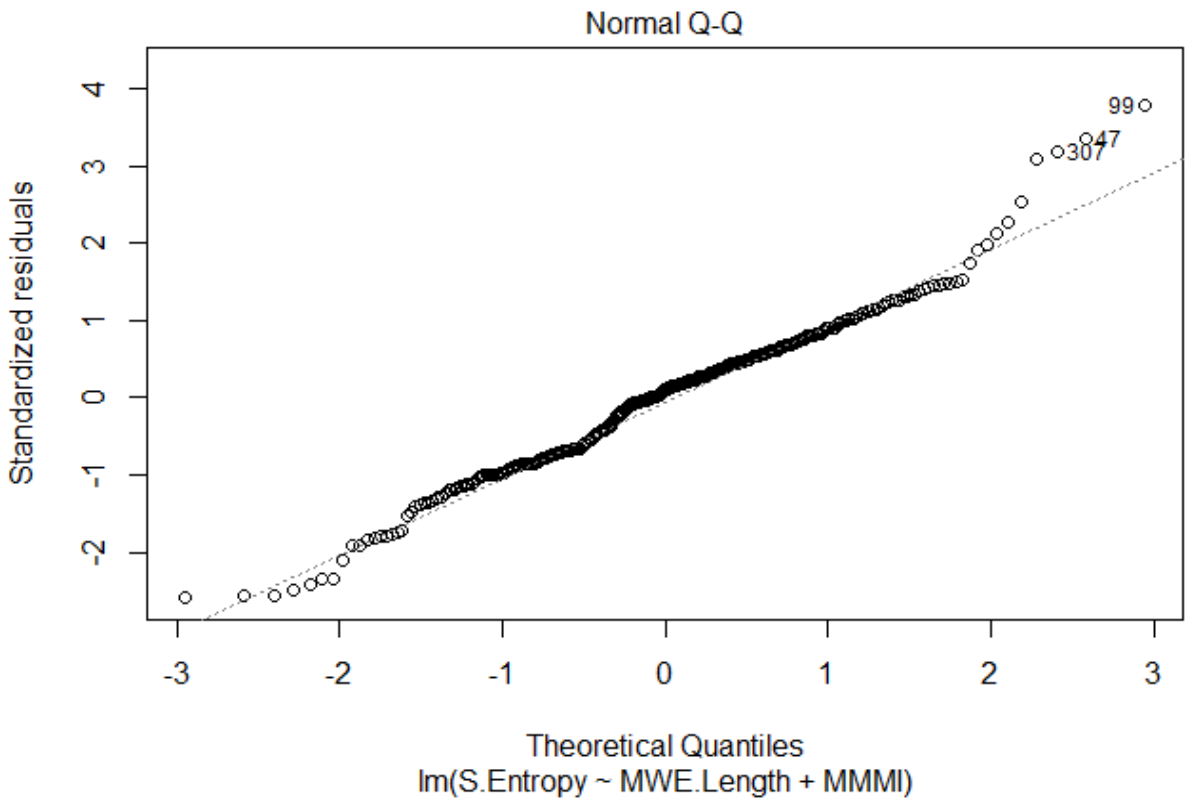
Because the *p*-value is less than 0.05, but "the D-W statistic is very close to 2," it is *equivocal* to say whether autocorrelation exists. But one thing needs to be pointed out is that MMMI is based on a first-order Markov chain, a kind of discrete stochastic processes. This can *inherently* contribute to autocorrelation.

Normality of Residuals

To check whether "there are any violations of normality", the Shapiro-Wilk normality test can be used:

```
shapiro.test(residuals(mBasics))

Shapiro-Wilk normality test
data:  residuals(mBasics)
W = 0.98277, p-value = 0.0008448

plot(mBasics, which = 2)
```

Since *p*-value is 0.0008448, which is less than 0.05, "the null hypothesis of normality" can be rejected. On the other hand, the normality of residuals is approximately upheld, which is reflected by the 'almost straight' line in Normal Quantile-Quantile graph. Thus, the evidence is inconclusive.



In summary, since some of the assumptions are only marginally met or even marginally violated, there is a need to explore another modeling technique to remedy the situation. And it is suggested that such a modeling technique is tentatively semiparametric regression, a middle ground between parametric (e.g. multiple linear regression) and nonparametric regression. It is provisional because semiparametric regression is new to quantitative linguistics and its properties have to be further explored and understood.

Empirical Findings on Research Questions and Hypotheses

The research questions and hypotheses formulated in section Research Methodology with Literature Review are recapped below and the hypotheses are empirically tested as follows:

1. Are Shannon Entropy and MWE Length empirically positively related?

   H_0: Shannon Entropy and MWE Length are <u>not</u> statistically positively related.

   H_1: Shannon Entropy and MWE Length are statistically positively related.

   The empirical findings in the previous sub-section Empirical Multiple Linear Regression *reject* the H_0 because:

   a) t-value = 21.76, p-value < $2e^{-16}$, which is less than 0.05

   b) Size of coefficient estimate is 0.107519, which is significant and positive

2. Are Shannon Entropy and Markov multivariate mutual information (MMMI) empirically inversely related?

   H_0: Shannon entropy and MMMI are <u>not</u> statistically inversely related.

   H_1: Shannon entropy and MMMI are statistically inversely related.

   The empirical findings in the previous sub-section Empirical Multiple Linear Regression *reject* the H_0 because:

   a) t-value = -10.42, p-value < $2e^{-16}$, which is less than 0.05

   b) Size of coefficient estimate is -0.023458, which is significant and negative

In short, these two cases of hypothesis testing should be interpreted with the concern that some of the regression assumptions are only marginally met and even marginally violated. The plausible implication is that the statistics might not be reliable.

Future Research

This section comprises three sub-sections of extensions, namely statistical idiomaticity, formulaic idiomaticity, and regression modeling. The first two are related to the theme concgram-based multi-word expressions (MWEs) in one part of the topic of the present research, while the third one is related to the empirical regression of entropy and multivariate mutual information (MMI) in another part of this topic. The whole topic can therefore be *extended* with a *tentative* small change in wording, but this does not imply a repeat of most of the contents of the present research:

Empirical Semiparametric Regression of Entropy & Multivariate Mutual Information
of Concgram-based Multi-Word Expressions on Learner Corpus

Statistical Idiomaticity

In section Linguistic Phenomena with Literature Review, the conception statistical idiomaticity is briefly mentioned. And it is recapped here: Baldwin & Kim (2010 pp. 269) contend that such "idiomaticity occurs when a *particular combination* of words occurs with *markedly high frequency*, relative to the component words or alternative phrasings of the same expression".

Fang's first insight is that the inverse relation between Shannon entropy and MMMI is found to be statistically significant in the present research and this is potentially interesting in the sense that shorter concgram-based MWEs are perhaps less determinate than longer ones, both in terms of their semantic meanings and in terms of their collocational behavior, something that are worth of discovery in future research. How can such a *cline* of determinacy and indeterminacy be statistically quantified?

Fang's another insight is that semantically and structurally, intervening words should be deemed to be replaceable. How can replaceability be *statistically* quantified? This points to another interesting issue: within a longer MWE construct, can there be a set of 'comment constructs' that demonstrate varying statistical significance? The one with a lower significance should, in this situation, be treated as a negligible part, i.e. intervening words.

Zooming in on MMMI, the prospective future research should ascertain: is it valid that the non-contiguous or hybrid concgram *cannot* be modeled by a first-order Markov chain, which *only* applies to the n-gram in the literature, especially Wei & Li (2013)? It is a big question because a first-order Markov chain is an integral part of the formulation of MMMI; if the answer is 'cannot', all MMMI values are across-the-board invalid. In the present research, it is expediently assumed that non-contiguous or hybrid concgrams are treated as if they were contiguous concgrams. If this

assumption collapses, reformulating the integration of a first-order Markov chain and/or pseudo-bigrams may be necessary.

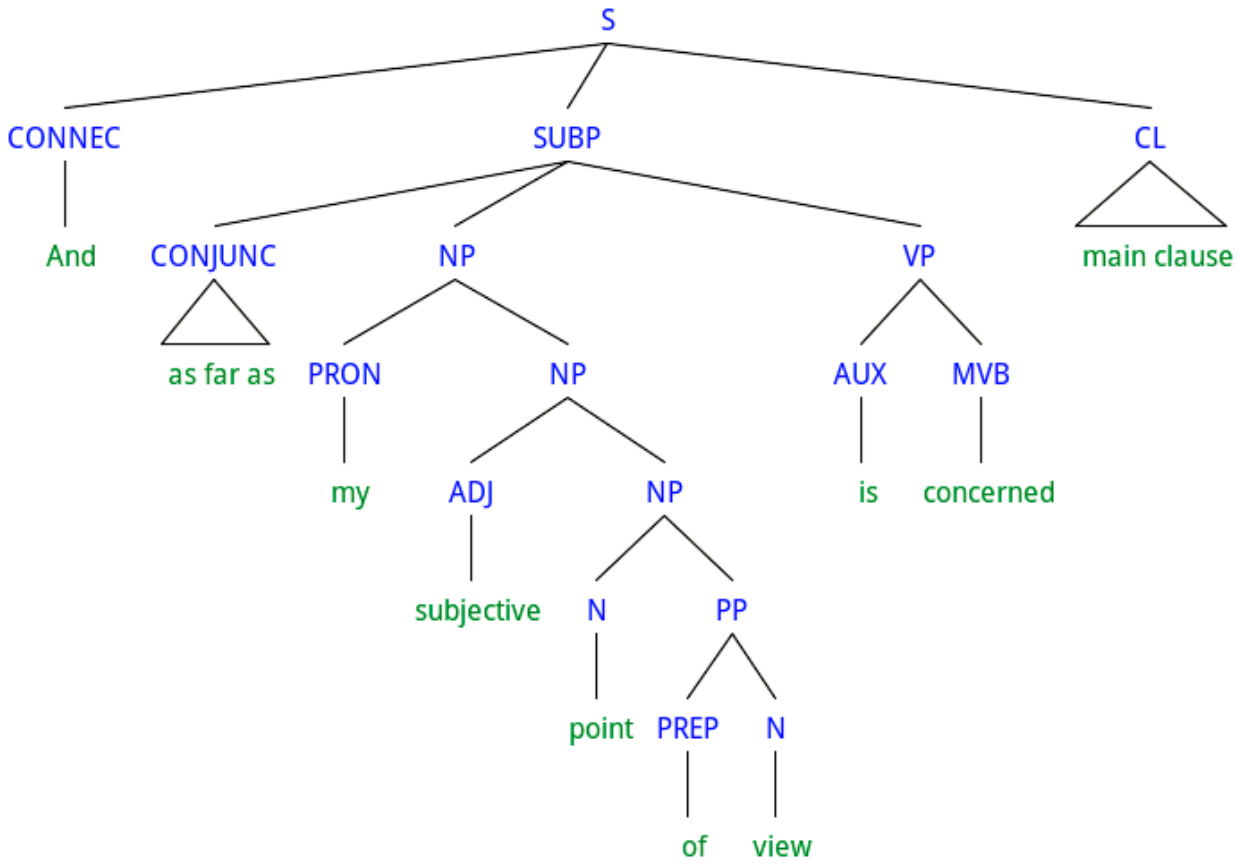
Formulaic Idiomaticity

To what extent is an MWE formulaic in the sense that it cannot be *structurally* dissected? As to the concgram "as far as…concerned", the end result of tagging and parsing the incomplete sentences that contain it by tagger AUTASYS and Alex's (or Survey) Parser conveyed by Fang (2007) is epitomized by the following one among the 67 instances of this concgram in the International Corpus of Learner English (ICLE) listed in section Authenticity of Evidence:

"And as far as my subjective point of view is concerned,"

As authoritative *Merriam-Webster Dictionary* defines, 'as far as' is a *conjunction* which means "to the extent or degree that". This is consistent with the parsing result shown in the Introduction.

MWE "as far as…concerned", an epitome of hybrid MWEs, is presented below by a *syntax tree* using software RSyntaxTree developed by Hasebe (2016). Such tree representation is not a usual practice in researching MWEs. But the present study trials it to see whether this anatomy of hybrid MWE "as far as…concerned" makes sense—can constituencies formulaically hang together be structurally dissected?



Syntax is incorporated into the concgram-based MWE, shown in tree above: on the tree, AVP means 'adverbial phrase', CONJ 'conjunction', NP 'noun phrase', VP 'verb phrase', ADV 'adverb', AUX 'auxiliary verb', and MV 'main verb'. The sub-tree NP is in the middle of the whole tree, in between sub-tree CONJ on the left and sub-tree VP on the right. Is this sub-tree NP *intervening words* that can be flexibly replaceable within the formulaic hybrid MWE "as far as…concerned"?

Regression Modeling

As sub-section Verifying the Regression Assumptions with R reveals, some of the assumptions of the multiple linear regression model are *only marginally met or even marginally violated* so there is a need to seek another method that can remedy such a situation as the estimates of the parameters in this model could be unreliable. The implication is that the findings might have been otherwise. And semiparametric regression as an apt option is a proposal rather than an insistence. So far semiparametric regression has not yet been applied to linguistics. Thus, there is ample room for an innovative solution to the above issue. But what does 'semiparametric' mean? In short, it is the *middle ground* among the trio 'nonparametric'—'semiparametric'—'parametric'. To understand semiparametric regression, some key concepts in statistics must be comprehended first and then notions of 'parametric' and 'nonparametric' regression.

Non-Rigorous Key Concepts of Statistics

As per en.oxforddictionaries.com/definition/, 'regression' means "a measure of the *relation* between the *mean value* of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost)". And 'parametric' means "assuming the value of a parameter for the purpose of analysis", whereas 'nonparametric' means "not involving any assumptions as to the form or parameters of a frequency distribution". Then, readers may be curious to know what does 'parameter' really mean since it is mentioned in these two definitions. 'Parameter' means "a numerical characteristic of a population, as distinct from a statistic of a sample", for example *mean of population* and *standard deviation of population*. 'Population' is defined as "a finite or infinite collection of items under consideration", while 'sample' is defined as "a portion drawn from a population, the study of which is intended to lead to statistical estimates of the attributes of the whole population." This series of definitions is meant to be as accessible as possible so as to arouse readers' interest.

Non-Rigorous Introduction to Regression Models

After presenting the non-technical exposition of key statistical concepts above, this section Future Research proceeds to enunciate a non-rigorous introduction to *regression models* such that the technicality in prospective future research can be kept as minimum as possible since it is linguistic rather than statistical research. Harnessing statistics is viewed from an *applied* point of view.

According to Mahmoud (2014), parametric regression, e.g. multiple linear regression in the present research, is a statistical model in which the function that describes the relationship between the response and explanatory variables is linear and known. "When the relationship between the response and explanatory variables is known, parametric regression models should be used." But "in many situations, that relationship is not known." Thus, in some situations, researchers need to "incorporate unknown, flexible, and nonlinear relationships between variables into their regression analyses", i.e. nonparametric regression, which "differs from parametric regression in that the shape of the functional relationships", i.e. nonlinear one, "between the response (dependent) and the explanatory (independent) variables [is] *not predetermined* but can be adjusted to capture unusual or unexpected features of the data." In brief, "if the relationship is unknown and nonlinear, nonparametric regression models should be used. But if the relationship between the response and *part of explanatory variables* is known while the relationship between the response and *the other part of explanatory variables* is unknown, semiparametric regression is worthy of trial.

Parametric and Nonparametric Regression

Fox (2016) offers relatively formal but still digestible exposition for wide readership by using many *vivid diagrams to visualize abstract concepts*. As computation is done by R, possibly intimidating statistical derivations can be hidden from linguists and other readers. Likewise, Crawley (2015) and Levshina (2015) explain how to deploy R mainly for applied *parametric* regression and briefly for *nonparametric* regression.

For detailed elucidation of nonparametric regression using R, Fox & Weisberg (2010) offer a substantial online appendix. And they pinpoint that "in traditional parametric regression models, the functional form of the model is specified before the model is fitted to data, and the object is to estimate the parameters of the model. In nonparametric regression, in contrast, the object is to estimate the regression function directly without specifying its form explicitly." Detailed executable R codes for nonparametric regression are provided by Fox & Weisberg (ibid.), the online appendix. They "describe how to fit several kinds of nonparametric regression models in R, including scatterplot smoothers, where there is a single predictor; models for multiple regression; additive

regression models; and generalized nonparametric regression models that are analogs to generalized linear models".

Hazelton (2015) suggests that nonparametric regression is statistical modeling for describing the "trend between a response variable and one or more predictors. This approach differs from classical regression models in that it does *not rely on strong assumptions regarding the shape of the relationship between the variables*. Rather, the data are allowed to speak for themselves in determining the form of the fitted regression functions." "Nonparametric regression can also be combined with parametric models to form hybrid *semiparametric models*".

Semiparametric Regression

Semiparametric regression is a *synthesis* of both parametric and nonparametric regression. Keele (2008) offers a vivid exposition of what semiparametric regression is. But Keele (ibid.) emphasizes that "the techniques here are *not meant to be a replacement* for models such as linear or logistic regression, but instead are meant to *enhance how these models are used* … the techniques can enhance data analysis but they are not a panacea." Keele (ibid.) adds that "while *nonparametric* regression can flexibly estimate nonlinear functional forms, it *cannot* make a correlation into a causal effect, it *cannot* make up for an omitted variable, and it *cannot* prevent data mining." But such models help visualize data so as to render linguists "more sensitive to *patterns* in the data that are often obscured by simply reading" on-screen parameter estimates.

Ruppert, Wand, & Carroll (2003) zoom in on the synthesis "of classical parametric regression techniques and modern nonparametric regression techniques to develop useful models", i.e. semiparametric regression. Accordingly, it is crucial to first acquire "a *good grounding* in the principles of parametric regression before proceeding to" semiparametric regression, which is more complicated. "In particular, some of the *theoretical aspects* of regression should be well understood since these are important in extensions to semiparametric regression."

Perhaps it is now the right timing to give more rigorous explication of semiparametric regression. As per Horowitz's (2015) exposition, "much empirical research in the social sciences is concerned with estimating *conditional mean functions*," i.e. $E(Y|x)$, which denotes taking the expected value or mean of a "dependent variable Y conditional on a vector of explanatory variables X." The most popular estimation methods presume that a conditional mean function "is known up to a set of constant parameters" estimated from data such that these methods are characterized as *parametric*. "Their use greatly simplifies estimation and inference but is rarely justified by theoretical or other a priori considerations." Statistical estimation and inference that thrive on convenient but unwarranted presumptions about "the form of the conditional mean function" may be enormously

unreliable. *Semiparametric* regression models can undermine "the strength of the assumptions required for estimation and inference, thereby reducing" the chances of obtaining unreliable results. Moreover, semiparametric methods can ameliorate certain disadvantages of nonparametric methods which completely "make no assumptions about the shape of the conditional mean function."

No matter parametric regression, nonparametric regression or semiparametric regression can be complex, depending on how they are presented technically. This sub-section adumbrates a *concise overview* of non-rigorous fundamentals of key statistical definitions, parametric and nonparametric regression, and semiparametric regression. All the relevant technical concepts and methods will be *sufficiently*, not rigorously, elucidated to facilitate a prospective *corpus and empirical* research into semiparametric regression if feasible. This sub-section brings out the essence of semiparametric regression for potential *practical* applications to linguistics.

So far there is only explication of statistics and no research question under section Regression Modeling. But all the elucidation is a preparation for understanding semiparametric regression. Can this type of regression really be *more appropriate* than parametric (e.g. multiple linear) regression used in the present research for modeling the relations between Shannon entropy and Markov multivariate mutual information (MMMI) as well as between Shannon entropy and MWE length in a prospective research? 'More appropriate' is a deliberately flexible wording such that there is room for fine-tuning it when better understanding of semiparametric regression is acquired in future research.

"Any application area that uses regression analysis can potentially benefit from" semiparametric regression. The present research just *preliminarily* explores semiparametric regression and *suggests* adopting it because once it is *acquired*, it will pay off in the long term.

Conclusion

The present research concludes that adopting an overarching bottom-up approach to the present research is a correct decision since vague research questions and non-testable hypotheses contrived without empirical grounds at the outset are avoided. Importantly, two findings that would not have been anticipated are obtained.

The first finding is that independent, explanatory, or predictive variable Markov multivariate mutual information (MMMI) tends to *negatively* relate to dependent variable Shannon Entropy such that the larger the MMMI (i.e. more information), the smaller the Shannon entropy (i.e. uncertainty). The second finding is that independent variable multi-word expression (MWE) length tends to *positively* relate to Shannon entropy such that the larger the MWE Length, the larger the uncertainty too. These two quantitative relations are visualized in a not only nice-looking but also revealing 3-D scatterplot using R.

But the analytics, especially Fang's insightful comment on the first finding, should go further: this finding can be appealing "in the sense that shorter concgrams are perhaps less determinate than longer multiword expressions, both in terms of their semantic meanings and in terms of their collocational behavior, something that are worth looking into". In what ways can such a *cline* of determinacy and indeterminacy be statistically quantified? Fang's another insight is "semantically and structurally, intervening words should be deemed to be replaceable, which should be shown *statistically*." This enlivens another compelling issue: whether within a longer MWE construct, can there be a set of '*comment constructs*' that demonstrates varying degree of statistical significance? The one that is insignificant should, in this situation, be treated as a negligible part, i.e. *intervening words*.

Readers may discern a superficial contradiction between the first and the second findings: the first one is a negative relation, whereas the second is a positive one. But it is necessary to point out although both findings involve the dependent variable Shannon Entropy, independent variables MMMI and MWE Length are *not directly comparable*. In fact, MMMI is a quantitative, continuous, ratio variable obtained by sophisticated derivation based on a first-order Markov chain and pseudo-bigrams (see Introduction), whereas MWE Length is a quantitative, discrete, interval variable with simple values 2, 3, 4 and 5 only (see Introduction).

Besides, is it valid that the MMMI of the non-contiguous or hybrid concgram cannot be modeled by a first-order Markov chain, which is so far applicable to n-grams or contiguous concgram-based MWEs only? For example, in "as far as my subjective point of view is concerned", the pattern is 'as far as…concerned', while 'my subjective point of view is' are *presumptively* intervening words,

which cannot weaken the *dominance* of the pattern in the formation of non-contiguous or hybrid concgram.

Moreover, to what extent is an MWE formulaic in the sense that it cannot be *structurally* dissected? The present study trials a conventional syntax tree representation, though not a usual practice in researching MWEs, to see whether this anatomy of hybrid MWE "as far as…concerned" makes sense—can constituencies formulaically hang together be structurally dissected? As for the above hybrid MWE, is the set of *presumptively* intervening words flexibly replaceable within the formulaic construct, i.e. pattern? For the parsing result, see Introduction; for the syntax tree, see sub-section Formulaic Idiomaticity.

And the multiple linear regression, a kind of parametric regression, built and tested in the present research *only marginally fulfills or even marginally violates* some of the assumptions of parametric regression. This can affect the reliability of the quantitative relations (findings) between Shannon entropy and MMMI as well as between Shannon entropy and MWE length. The implication is that the findings might have been otherwise. Thus, it is necessary to seek another remedial methodology. Is semiparametric regression, a synthesis of parametric and nonparametric regression, an appropriate option for modeling those two relations?

Before explicating this potential alternative solution, it is crucial to accessibly adumbrate what 'parametric' and 'nonparametric' mean in a non-rigorous way as they constitute the conception 'semiparametric' and their definitions are meant to be as digestible as possible so as to *arouse readers' interest*. And 'parametric' means "assuming the value of a parameter for the purpose of analysis", whereas 'nonparametric' means "not involving any assumptions as to the form or parameters of a frequency distribution". Existing in these two definitions, 'parameter' means "a numerical characteristic of a population, e.g. *mean of population* and *standard deviation of population*," unlike a statistic of a sample.

It is advisable to use parametric regression when the relationship between the dependent and independent variables is known. But such a relationship is often unknown. Accordingly, it is suggested that nonparametric regression should be used. Still, if the relationship between the dependent and *part of explanatory variables* is known while that between the response and *the other part of explanatory variables* are unknown, semiparametric regression should be trialed.

Acquiring semiparametric regression can be not so straightforward but once it is *acquired*, it will pay off in the long term. In sum, it is worth investing effort in its acquisition. Noteworthily, it is suggested that the techniques of semiparametric regression are *not meant to replace* models "such as linear or logistic regression, but instead are meant to *enhance how these models are used* … the techniques can enhance data analysis but they are not a panacea."

After the non-rigorous exposition above, if readers preliminarily thrill to these three types of regression, among which semiparametric regression is the middle ground, prospective future research lets readers comprehend relatively technical references even though the rigorousness is kept as minimum as possible as it is linguistic rather than statistical research. More precisely, linguistics is the end while statistics is the means. Harnessing statistics is viewed from an *applied rather than theoretical* point of view. Indeed, as statistical computation is done by R, possibly intimidating statistical derivations can be hidden from linguists and readers. But theories cannot be completely avoided.

So far semiparametric regression has not yet been applied to linguistics. Thus, there is ample room for an innovative solution to the issues discussed above. Because semiparametric regression is a synthesis of parametric and nonparametric regression, all results obtained in the present research by (parametric) multiple linear regression can be used as essential empirical references for modeling semiparametric regression of Shannon entropy, MWE length and Markov multivariate mutual information (MMMI) in prospective research. (see sub-section Semiparametric Regression for detailed explanations)

(Grand Total: 18,064 words, excluding Table of Contents, Acknowledgement and Bibliography)

Bibliography

Abu-Ssaydeh, A.-F. (2006). Multi-word units: Can lexicography have a role in their acquisition? In *Babel*, 52: 4, 349–371.

Baclawski, K. (2008). *Introduction to probability with R*. CRC Press.

Baldwin, T. & Kim, S.N. (2010). Multiword expressions. In N. Indurkhya & F.J. Damerau (Eds.) *Handbook of natural language processing*. CRC Press.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. In *International Journal of Corpus Linguistics*, 14(3), 275-311.

Biber, D. (2012). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine and H. Narrog (Eds.) *The Oxford handbook of linguistic analysis*. Oxford University Press.

Cheng, W., Greaves, C., Sinclair, J.M. & Warren, M. (2008). Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams. In *Applied Linguistics*.

Cheng, W., Greaves, C. & Warren, M. (2006). From n-gram to skipgram to concgram. In *International Journal of Corpus Linguistics*, 11:4, pp. 411-433.

Cheng, W., & Leung, M. (2012). Exploring phraseological variations by concgramming: The realization of complete patterns of variations. In *Linguistic Research*, 29(3), 617-638.

Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, Vol. 16, No. 1.

Conklin, K. & Schmitt, N. (2012). The processing of formulaic language. In *Annual Review of Applied Linguistics*, 32, 45–61.

Crawley, M.J. (2015). *Statistics: an introduction using R*. Wiley.

da Silva, J.F. & Lopes, G.P. (1999). A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora. In *Proceedings of the 6th Meeting on the Mathematics of Language*, pp. 369-381. http://hlt.di.fct.unl.pt/jfs/MOL99.pdf

Ebeling, S.O. & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin & F. Meunier (Eds.) *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Ellis, N.C., Simpson-Vlach, R., Römer, U., O'Donnell, B.M. & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition research. In S. Granger, G. Gilquin & F. Meunier (Eds.) *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. In *Text*, 20(1), pp. 29-62.

Fang, A.C. (2007). *English corpora and automated grammatical analysis*. Commercial Press.

Faraway, J.J. (2015). *Linear models with R*. CRC Press.

Feng, Z. & Hu, F. (2012). *Shuli yuyanxue* [数理语言学]. Shangwu yinshuguan.

Fox, J. (2016). *Applied regression analysis and generalized linear models*. SAGE. (downloadable datasets at www.sagepub.com/fox3e).

Fox, J. & Weisberg, S. (2010). *An appendix to an R companion to applied regression*. SAGE. https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Nonparametric-Regression.pdf

Fox, J. & Weisberg, S. (2011). *An R companion to applied regression*. SAGE.

Geman, S. & Johnson, M. (2003). *Probability and statistics in computational linguistics*. www.dam.brown.edu/people/geman/Homepage/Computational%20linguistics/Review_IMA.pdf

Ghahramani, S. (2016). *Fundamentals of probability: with stochastic processes*. CRC Press.

Gilquin, G. & Granger, S. (2015). Learner language. In D. Biber & R. Reppen (Eds.) *The Cambridge Handbook of English corpus linguistics*. Cambridge University Press.

Goldsmith, J. (2003). *Probability for linguists*. Department of Computer Science, University of Chicago. http://people.cs.uchicago.edu/~jagoldsm/Papers/probability.pdf

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.) (2009). *International corpus of learner English*. (with CD-ROM). Presses universitaires de Louvain.

Gray, B. and Biber, D. (2015). Phraseology. In D. Biber & R. Reppen (Eds.) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press.

Greaves, C. (2009). *ConcGram© 1.0: a phraseological search engine*. John Benjamins.

Greaves, C. & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O'Keeffe & M. McCarthy (Eds.) *The Routledge handbook of corpus linguistics*. Routledge.

Gries, S.T. (2013). Statistical tests for the analysis of learner corpus data. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.) *Automatic treatment and analysis of learner corpus data*. John Benjamins.

Gries, S.T. (2015a). Statistics for learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.) *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Gries, S.T. (2015b). Quantitative linguistics. In J.D. Wright (Ed.) *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., Vol. 19.

Gries, S.T. (2015c). Quantitative designs and statistical techniques. In D. Biber & R. Reppen (Eds.) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge University Press.

Gries, S.T. & Deshors, S.C. (2015d). EFL and/vs. ESL?: A multi-level regression modeling perspective on bridging the paradigm gap. In *International Journal of Learner Corpus Research*, 1:1, pp. 130-159.

Gries, S.T. (2017). *Quantitative corpus linguistics with R: a Practical Introduction*. Routledge.

Hasebe, Y. (2016). *RSyntaxTree*. http://yohasebe.com/rsyntaxtree/

Hazelton, M.L. (2015). Nonparametric regression. In *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., Vol. 16.

Horowitz, J.L. (2015). Semiparametric models. In *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., Vol. 21.

Jurafsky, D. & Martin, J.H. (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Keele, L. (2008). *Semiparametric regression for the social sciences*. Wiley.

Kumova Metin, S. & Karaoğlan, B. (2011). Measuring collocation tendency of words. In *Journal of Quantitative Linguistics*, 18:2, 174-187.

Levshina, N. (2015). *How to do linguistics with R: data exploration and statistical analysis*. John Benjamins.

Mahmoud, H.F.F. (2014). *Parametric versus semi/nonparametric regression models*, Virginia Polytechnic Institute and State University, http://www.lisa.stat.vt.edu/?q=node/7517.

Martinez, R. & Murphy, V.A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. In *TESOL Quarterly*, Vol. 45, No. 2, June.

Nissim, M. & Zaninello, A. (2013). Modeling the internal variability of multiword expressions through a pattern-based method. In *ACM Transactions on Speech and Language Processing*, Vol. 10, No. 2, Article 7, June.

Paquot, M. & Granger, S. (2012). Formulaic language in learner corpora. In *Annual Review of Applied Linguistics*, 32, 130-149.

Paquot, M. & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. In https://www.researchgate.net/publication/313108575

Ruppert, D., Wand, M.P., & Carroll, R.J. (2003). *Semiparametric regression*. Cambridge University Press.

Smart Words (2013). *Linking words*. www.smart-words.org/linking-words/linking-words.pdf

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Ugarte, M.D., Militino, A.F. & Arnholt, A.T. (2016). *Probability and statistics with R*. CRC Press.

Unwin, A. (2015). *Graphical data analysis with R*. CRC Press.

Van de Cruys, T. (2011). Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pp. 16-20.

Wei, N. & Li, J. (2013). A new computing method for extracting contiguous phraseological sequences from academic text corpora. In *International Journal of Corpus Linguistics*, 18:4, 506–535.

(Total: 54 references)