Building ESP Corpus-driven Specialized Words in Vocational Education in Saudi Arabia

Waleed Mahmoud Hamdoun
Whamdoun77@gmail.com

doi: https://doi.org/10.37745/bjmas.2022.04151 Published August 05, 2024

Citation: Waleed Mahmoud Hamdoun (2024) Building ESP Corpus-driven Specialized Words in Vocational Education in Saudi Arabia, *British Journal of Multidisciplinary and Advanced Studies* 5(4),12-43

ABSTRACT: *Both corpus revolution and advances in computer-based technologies have a remarkable impact on teaching and learning English as well as corpus studies. Nevertheless, Jablonkai (2020) observed that professional or occupational programs are less studied or researched compared to academic focus courses. Therefore, this paper is based on a replication study of Coxhead and Demecheleer (2018, cited in Jablonkai and Csomay, 2023) to develop ESP corpus-driven specialized words in vocational education. This process is tackled in three main phases: collection, computerization, and annotation of data. Data collection will be in mixed method from written and spoken ESP corpuses before using qualitative and quantitative analysis to measure frequency, key words and concordance comparing with other ESP corpora. The computerization phase will be based on lexical frame analysis (LFA) framework to convert a text to corpus, and this is quite labor-intensive part of creating the corpus. Finally, the annotation phase includes further lexical information covering semantic, pragmatic, and syntactic details. The first phase, data collection and analysis, will be reflected in Coxhead and Demecheleer's (2018) and the remaining two phases, annotating and computerizing data, will be elaborated through using LFA with two similar Business English corpora, PROLEX and PROCOMPARE, cited in Vincent (2009). The findings and conclusion of using three corpora could be effective in conducting this study as a preparation for creating a comprehensive technical corpus covering specialized words for all oil industry majors in my context. Further research could be conducted in the service of replication studies and expanding areas of ESP and vocabulary research using corpora.*

KEY WORDS*:* lexical acquisition, replication study, annotation, tagging, specialized word, vocational education

INTRODUCTION

According to recent studies, the growth of corpus approaches is clearly related to advances in computer-based technologies**.** The early generation of researchers interested in building corpus had to scan material by hand and then run OCR software over the scans, but current researchers can use powerful software to batch-convert PDF documents into text documents for inclusion in a corpus**.** Alongside this, the availability of freeware programs enable both teachers and learners to become researchers of their own texts as well as those of others**.** To confirm this perception, Jablonkai and Csomay (2023) point out that " the corpus “revolution” has had a powerful impact on English language teaching and corpus studies which have been influential in the field of ESP**.** Furthermore, Jablonkai and Csomay (2023, P. 206) quoted Anthony's (2018) definition of ESP as *"an approach to language teaching that targets the current and/or future academic or occupational needs of learners, focuses on the necessary language, genres, and skills to address these needs*".

To develop actual materials for academic or occupational learners, Corpora can help identify words that appear more frequently in one field than another, For example, in general English vs in ESP. Further, Woodrow (2018) points out that "*the advantages of using corpora are that real, authentic usage of language can be uncovered"* and therefore a language corpus is a source of evidence about authentic language use However, it is believed that a general EAP collocation list would not suffice to support students in ESP and discipline-specific collocation resources need to be developed to meet the authentic needs of the target learners. Accordingly, Parkinson & Mackay (2016) conclude that learning and using the specialized vocabulary of a subject or profession is important for identity and community belonging as added by Wray (2002).

In short, understanding specialized vocabulary in texts is important for some reasons: it defines the amount of learning that is required for ESP learners and the need for planning for this lexis in ESP courses. To this end, my study is aimed at using texts of corpora as well as spoken corpora related to the vocational education in oil industry training in the service of support my Saudi context trainees. The following parts will elaborate how to use corpus to derive specialized words using corpus tools, especially lexical frame analysis aspects. Further, Laurence Anthony's (2019a) AntCorgen software allows users to create their own discipline specific corpora using the internet and carry out their own lexical analysis using other tools such as AntConc. In addition to this, this paper elaborates the framework of lexical analysis of specialized words driven from an ESP corpus based.

To end this part, let me express my agreement with Jablonkai's (2020) observation that there is "*an imbalance in ESP research between professional and academic focus*" confirming that ESP or vocational programs are less studied. For this reason, I am really interested in conducting this study to develop a corpus-derived wordlist resource enriched in collocational information for oil industry training organizations in Saudi Arabia. This corpus should be built from text samples identified as relevant to oil industry education in vocational training institutions through conducting a survey for all technical department staff, around 120 people including male and female from different countries around the world in line with interviews of a group of 10-12 technical instructors, 1-2 instructors from each major, at Energy Tech College in Dammam City in Saudi Arabia.

Study Aims and Research Questions

As mentioned in the preceding paragraphs, the main purpose of this study is to develop a corpus-derived wordlist resource enriched in collocational information for oil industry training organizations in Saudi Arabia through achieving the following study aims:

- To identify specialized technical words in oil industry majors in vocational education using corpus-based research.

- To apply the LFA framework to two sublanguage corpora for the acquisition of lexical information.

-To consider what corpora can provide for the language engineer.

-to provide a syntactically sophisticated computer-processed frequency-count of professional lexis with systematic indications.

- to classify text samples according to those typological categories deemed to be significant for oil industry vocational training.

- to discover the collocational range of those lexemes constituting the professional or vocational lexis.

This paper is trying to answer four research questions:

1- To what extent corpora can help identify words that appear more frequently in one field than another.
2- To what extent 'corpora are more interesting than dictionaries as a source of linguistic knowledge,
3- To what extent ESP corpus-based research can identify specialized word lists.
4- To what extent corpora are useful for the language engineer.

Study Significance

This research is based on my own interest in investigating language used in oil industry education. One aim of this study is to find out more about specialized vocabulary in the field of oil industry through interviews with the technical instructors of the technical department at Energy Tech college, my workplace in Saudi Arabia to survey and interview representatives for majors at my Saudi context: pipefitting, operation, scaffolding, rigging, welding, drilling, and crane operation. One of the key concerns is the need to build knowledge of this lexis with all technical trainees according to the preceding specifications. Additionally, one purpose of my study is to develop a specialized word list for each major of oil industry training to support our trainees in their vocational training. The collocations for each target major should be extracted from around 10% of texts in the ESP corpus using the criteria of a mutual information (MI) score above 4, and a minimum frequency above 5.

Study Limitations

I totally agree with Miller and Biber's (2015, cited in Jablonkai and Csomay, 2023) assumption that it could be difficult to replicate findings of corpus-based studies, even when applying the same principles. Their principal concerns include whether a corpus represents the academic or specialized domain and corpus design and content in relation to linguistic variation, and I think this may affect the credibility and reliability of the study. The second issue is related to the expanding areas of specialized vocabulary and ESP research using corpora. As noted above, EAP areas are well serviced in corpus-based research, while some areas of ESP remain relatively untouched, especially vocational education. This requires a deeper understanding of vocabulary in ESP and, therefore, more support for ESP learners and instructors.

LITERATURE REVIEW

The following parts of this section will review the existing literature in terms of corpus and lexicon, lexical acquisition types, lexical analysis, to computerize sample texts to ESP corpus based.

Linguistic evidence: from text to corpus

Krishnamurthy (2008) agreed with John Sinclair's perception that the use of natural language is considered the best source of linguistic evidence. For this reason, a lexicography serves as a commentary on examples that are selected automatically for their typicality from authentic texts to show more details about lemma or headword to have accurate interpretation to the word meaning and its form. Krishnamurthy (2008, p.236) concluded that "*A corpus-driven approach involves a bottom-up methodology, beginning by selecting unedited examples from the corpus, identifying their shared and individual features, and only then grouping them for the purpose of lexicographic presentation."* Further, Sinclair (1966, cited in Krishnamurthy 2008, P. 235) concludes that "*all linguistic study must start from text: 'Every morpheme in a text must be described both grammatically and lexically.'*

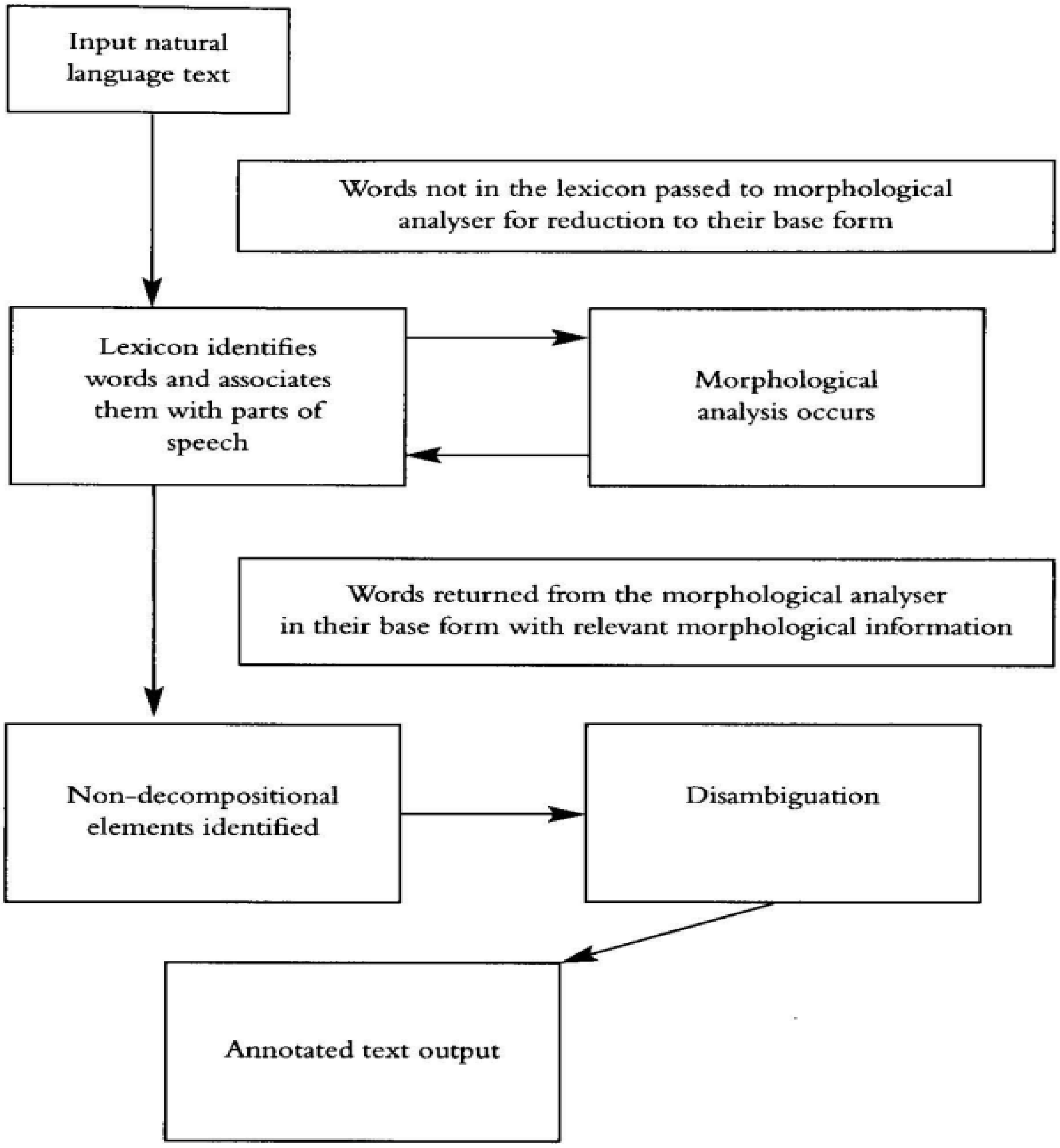
Corpora in Lexical Studies

McEnery and Wilson (2001) pointed out that Corpora currently play a great role in lexical studies, in the sense that the lexicographer can use the computer for a few seconds to call up all the examples of the usage of a word or phrase from many millions of words of a text. I totally agree with McEnery's perspective because it is useful not only for revising dictionaries very quickly, but also for adding up-to-date information about the language to make the target definitions complete and precise.

Corpora and Language Engineering

McEnery and Wilson (2001, 133) point out that *"language engineering is principally concerned with the construction of viable natural language processing systems for a wide range of tasks*." It is essentially a *'rather pragmatic approach to computerized language processing' which seeks to bypass the 'current inadequacies of theoretical computation linguistics'.*

Figure 1: A schematic design for a part-of-speech tagger, adapted from McEnery and Wilson (2001, p. 137)

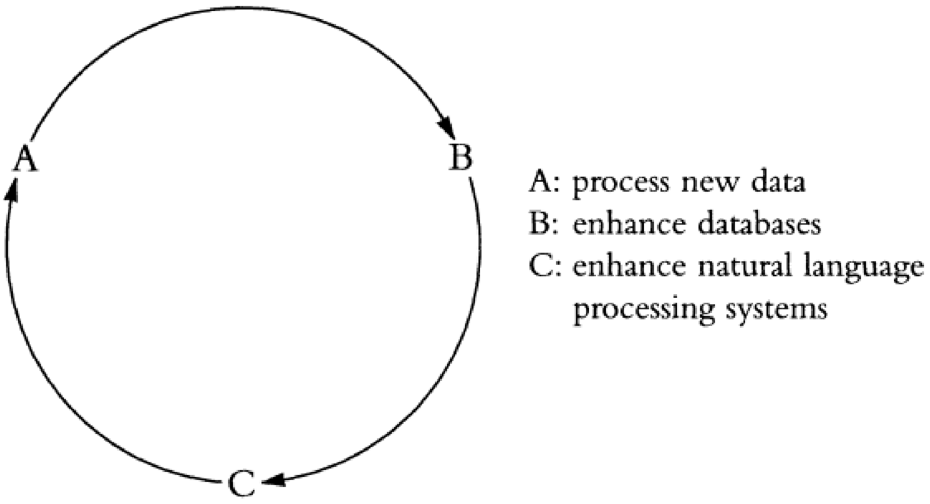

Input natural language text
Words not in the lexicon passed to morphological analyser for reduction to their base form
Lexicon identifies words and associates them with parts of speech
Morphological analysis occurs
Words returned from the morphological analyser in their base form with relevant morphological information
Non-decompositional elements identified
Disambiguation
Annotated text output

The following parts will examine the emerging concerns regarding how the computer can be made to derive a lexicon from real texts, including the question of which computational techniques emerge as most suitable for deriving the relevant types of lexical information.

The Relation between the Lexicon and the Corpus

According to Vincent (2009, p.67), there is no such direct relationship between the lexicon and the corpus, both entities are crafted for different purposes. Additionally, Vincent defines a lexical database saying that it is " *a computerized lexicon which is structured to record morphological, syntactic, semantic, pragmatic, and sometimes phonological information about the word".* It is crucial to interface the corpus with such a database to update the database from time to time, since the corpus is updated with new texts in the service of monitoring the state of the language or sublanguage from time to time. Additionally, figure 2 below shows the relationship between the linguistic database and the corpus according to Leech's (1987) conception. Where A represents corpus data which is then processed to create/ enhance B, the linguistic database, which in turn enhances the natural language processing (NLP) system, C, itself seeking to improve the system's performance by processing new data to enhance the database.

Figure 2: The relationship between a linguistic database and a corpus (Leech 1987, from *Vincent,2009, p.68*)



To sum up, I strongly agree with Leech's (1987, cited in Vincent, 2009) perception that *"the close relationship between a linguistic database and a corpus is seen especially in probabilistic systems where 'the frequency data derived from corpora are virtually indispensable for system updating and enhancement' of the database.*"

Also, Vincent asserts that *"the linguistic information enhancing such a database should come from various sources, of which machine-readable dictionaries and textual corpora are probably the most significant".'* (p.69)

Kim (1991, cited in Vincent 2009) points out that *" the lexical database can be viewed in terms of - to borrow a concept from knowledge-based systems - its conceptual and computational structures. A conceptual structure is a format suitable for humans to understand, whereas a computational or software structure is one more appropriate for computers.*" as shown in figure 3 below.

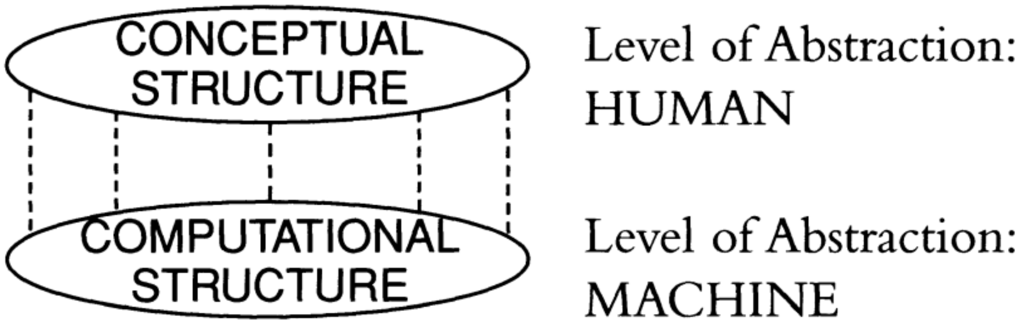


Figure 3: Representation phase of a lexical database, viewed from two abstractions: human and computer implementation (Vincent 2009:70).

Vincent (2009, p.70) adds that "*the computational structure is an explicit structure which directly reflects the conceptual structure, given the position that linguistic and conceptual structures are not dissimilar, unless proven otherwise"*. In turn, the conceptual structure of the lexical data/knowledge base assists in being a theoretical linguist's dictionary from which the lexicographer can utilize various knowledge structures to produce, for instance, a Dictionary of Advanced Learner's Business English.

Vincent (2009) assures the importance of databases in updating and maintaining dictionaries as well as enabling the exchange and sharing the information among projects, especially the automatic generation of several printed versions of a dictionary.

Lexical Acquisition

As explained by Vincent (2009), lexical acquisition may be regarded as a rubric for both 'lexical knowledge acquisition' and 'lexical data acquisition'. The two primary data for lexical acquisition include either MRDs or corpora, or a combination of both these main sources of lexical data. Additionally, he concluded that a corpus can be used as an on-line textual resource for lexical acquisition for two main assumptions: firstly, the corpus is relatively representative of the variety of (sub)language it has been gathered for. Secondly, the tools used for such acquisition are reliable.

Manual Lexical Acquisition

According to Nirenburg and Raskin (1987, cited in Vincent, 2009, P.74), the basic approach to lexicon building is that *'the work is done by humans assisted by an interactive aid which enhances productivity and ensures uniformity*' Further, lexicon building in natural language processing (NLP) involves the acquisition of three interrelated but distinct lexicons, which include the following:

> *the world concept lexicon which structures our knowledge of the world, the analysis lexicon which is indexed by natural language words and phrases connected with concepts from the world concept lexicon, and the generation lexicon, which is indexed by concepts in the world concept lexicon connected with natural language words and phrases.* (P.74)

Ingria et al. (1992, cited in Vincent, 2009, p.75) note that *'the lexicons associated with the CMT -CMU analysis and generation systems are among the largest and broadest available in NLP systems. This is due in part to the facility with which data is entered by the human knowledge engineer'*. In addition to this, Velardi and Pazienza (1991, p.15 7) asserts that *'a manual codification of the lexicon is a prohibitive task, regardless of the framework adopted for semantic knowledge representation; even when a large team of knowledge enterers is available, consistency and completeness are a major problem.*"

Automatic and Semi-automatic Lexical Acquisition

Vincent (2009) finds out that automatic and semi-automatic lexical acquisition could be illustrated in two dimensions; firstly, the two different methodologies of either (a) extracting it from raw text or (b) processing the text first. Secondly, the two different methodologies of using either (a) knowledge-based approaches or (b) statistical ones. However, there seems to be a tendency towards producing hybrid methodologies combining these methods.

According to Vincent (2009), the extraction of word associations from corpora has been the subject of several recent studies. Smadja (1989) points out that:

> *this focus on lexical co-occurrence knowledge has arisen because 'lexical relations' embody knowledge necessary for the proper usage of words ... and they represent the extent to which an item is specified by its collocational environment independently of syntactic or semantic reasons'. (p.76)*

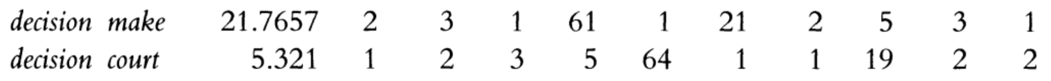
Furthermore, Smadja asserts that the acquisition of lexical co-occurrence knowledge in 'computational dictionaries' becomes very useful as this helps *'language generators correctly handle collocational restricted sentences.'* In other words, Calzolari and Bindi, 1990, cited in Vincent (2009) assume that when lexical collocations are supplied systematically in a computational lexicon, which is also annotated for frequency, which in turn make these collocations helpful for lexical disambiguation in analysis and crucial for lexical selection in generation.

the system XTRACT

Smadja (1989, 1991) adds that the system XTRACT is used to acquire collocational relations from the statistical analysis of large textual corpora, the extraction algorithm *'takes as input a corpus*, a span *parameter (five) and a dictionary specifying closed-class words,* especially articles, prepositions … etc. Martin (1983) demonstrated that 'more than 95% of all relevant lexical relations is obtained by examining collocates within a span of -5 and +5. Vincent (2009) illustrates that, for Martin (1983), the span is the 'co-text within which the collocates are said to occur', and the span position of a collocate is the number which specifies the distance of the collocate from the node. Additionally, the node refers to 'the lexical item whose collocational pattern we are looking for' and a collocate may be defined as 'any lexical item which co-occurs with the node within the specified co-text'. For example, in the idiomatic expression 'kick' the bucket', the collocate 'bucket' appears at span position + 2 of the nodes 'kick.

Nincent (2009) added that Smadja's algorithm produces a list of tuples (w1, w2, F), where (w1, w2) is a lexical relation between two open-class words (w1 and w2) identified in the corpus, and F is the frequency of appearance observed'. The frequency of common appearance of the two items is derived from 11 numbers representing the lexical relations in the corpus.

Let us consider the following sample output1 produced by XTRACT.



```
decision  make     21.7657   2    3    1   61    1   21    2    5    3    1
decision  court     5.321    1    2    3    5   64    1    1   19    2    2
```

The first figure is a factor which is empirically determined according to the size and nature of the corpus, and the other ten are the values which range from span position -5 to + 5. In the first example, 'decision make', the most common frequency numbers are '61' at -2 span position (i.e. make occurs to the left of the node decision and is separated by one word) and the second example, '64' at -1 span position (i.e. court occurs immediately to the left of the node decision), thus presumably making the most frequent expressions to be make (a/the) decision and court decision (compound noun) respectively.

the CIAULA and the ARIOSTO_LEX system

According to Vincent (2009), Basili et al. (1992a, 1992b, 1996):

> *offer a hybrid methodology which combines statistics with knowledge-based methods for lexical acquisition from corpora as demonstrated by both the CIAULA and the ARIOSTO_LEX system. The architecture and processing of the ARIOSTO_LEX system, a lexical learning system based on collocational analysis, consists of linearly going through the automated modules of morphologic[al] analysis, text segmentation, shallow syntactic analysis, semantic tagging (a phase which currently uses rapid human input), and clustered association.* (p. 79)

Since Basili et al. report good results with the hybrid which means a mix of statistics and linguistics/knowledge-based methodology adopted, such systems seem promising for the acquisition of lexical information from English corpora. The architecture for the lexical acquisition of English data would have to be modified. Further, Basili and co-workers assert their ability to reduce the size of the input corpus needed to generate reliable statistics. Generally, acquisition methods which rely on statistics cannot be applied by itself to small corpora since the statistics generated might not be reliable. To conclude, hybrid methodologies involving the right combination of statistics and rule-based approaches will not only be most practical, but also most successful in the acquisition of lexical knowledge.

The Lexicographer/Linguist's workbench for Lexical Acquisition

Jacobs (1989) warns that the strategy of building lexical knowledge bases by hand should not be disregarded, especially 'where the lexicon must include information that simply cannot be obtained otherwise'. It seems that lexicographic judgement must be exercised, and this includes treating both human intervention and fully automated methods as taking place against the background of a convergence between 'i) lexical and textual projects, ii) computational and traditional lexicography, and iii) statistical and rule-based approaches.'

Leech (1992b, cited in Vincent, 2009, P. 81) illustrates as four models concerning the role of corpus processing:

> ***Model A***, *the Linguistic Information Retrieval Model, is concerned with the processing of the corpus only to produce such lists as concordances to aid the human analyst in the study of language.*
>
> ***Model B,*** *the Induction Model, is concerned with the ability of the computer to induce generalizations from data.*
>
> ***Model C***, *the Automatic Corpus Processing Model, is concerned with the ability of the system to annotate unrestricted texts with such linguistic information as grammatical relations in as automatic manner as possible, since the system usually deals with thousands, even millions, of words of text.*
>
> ***Model D,*** *the Self-Organizing Model, the computer learns to train itself by progressively fitting the analysis to the data concerned, by using iterative re-estimation algorithms such as the Forward-Backward algorithm (Sharman 1989a, cited in Vencint, 2009, p.81*)

Zernik (1991, cited in Vencint 2009) points out the home truth that 'unfortunately, text is given in raw form', the automatic linguistic enrichment of the text might constitute a first step towards lexical acquisition. Once the processed text is available, the task of lexical acquisition then consists of providing mappings between the lexical units and the elements of the processed text. Accordingly, learning can be achieved from processed text, a first step towards extracting lexical information from a corpus is ideally to annotate the text with such levels as the following,

- **Word frequency counts and Concordance:** This represents the basic tool of lexicographers. Keyword-in-context and word-frequency profiles can be generated using such software.
- **Interactive searching:** To search and display patterns, especially word strings.
- **Lemmatization:** To relate a particular inflected form of a word to its base form or lemma, thereby enabling the production of frequency and distribution figures which are less sensitive to the incidence of surface strings
- **Word-Tag Extraction:** This is a more sophisticated version of word count, frequency, and sorting tools whereby a wordlist can be extracted from a preprocessed text file
- **Collocation:** To compute the statistical association of word forms in the text'
- **Part-of-speech-labelling-** To assign a word class or part-of-speech label to every word.
- **Syntactic Parsing:** To assign a fully labelled syntactic tree or bracketing of constituents to sentences of the corpora.
- **Semantic tagging, parsing, and sense disambiguation**: This tagged text is then fed into a semantic analysis program which assigns semantic tags representing the general sense field of words from a lexicon of single words and an idiom list of multi-word combinations (e.g. as a rule), which are updated as new texts are analyzed.
- **Pragmatic tagging:** is probably the least developed in terms of being automatically possible.
- **Link to lexical database:** to integrate the instances of words or phrases in a corpus with a structured lexical database!

Finally, Leech (1997, cited in Hunston 2022, P. 111) argued that "*annotation provided added value to a corpus by enabling the retrieval of information that was not available in other ways."*

Theoretical Framework: Applying LFA Framework to two Corpora of Business English

The main purpose of this study is to apply the LFA framework to two sublanguage corpora, the PROLEX corpus and the PROCOMPARE corpus, in the service of the acquisition of lexical information using a corpus-based approach.

The Notion of sublanguage, Genre, and Register

These two corpora- the PROLEX and the PROCOMPARE- are chosen for their sublanguage properties; both McNaught (1993) and Frawley (1988), cited in Vincent 2009, p. 123), note the following:

- *Sublanguage is strongly lexically based.*
- *Sublanguage texts focus on content.*
- *Lexical selection is syntactified in sublanguages,*
- *Surface collocation plays a major role in sublanguages.*
- *Sublanguages demonstrate elaborate lexical cohesion.*

Harris (1991, cited in Vincent 2009) points out that a sublanguage is a part of natural language with a grammar of its own: 'certain proper subsets of the sentences of a language may be closed under some or all the operations defined for the language. In other words, (Sager 1986) concludes that sublanguage refers to a language used by a 'particular community of speakers, for example, those concerned with a particular subject matter or those engaged in a specialized occupation'. Accordingly, a sublanguage is like linguistic notions such as register and genre.

Further, Halliday and Hasan 1985, cited in Vincent 2009, p. 124) define register as *"a variety of language according to use which creates contextual meaning through its three elements: field, tenor, and mode. These three aspects of the social context always act upon the language as it is being used."* Also, in terms of a functional approach to language, register is a *semantic concept* whereby *field, tenor, and mode* are correlated with the *ideational, interpersonal, and textual functions* respectively.

Halliday and Hasan (1985) pinpoint the notion of text structure Ruqaiya Hasan conceptualizes a Contextual Configuration (CC) as a specific set of values that realizes the field, mode, and tenor of discourse. The CC is used to predict the obligatory and optional elements of text structure to specify the following:

- What elements must occur.
- What elements can occur.
- Where must they occur.
- Where can they occur.
- How often can they occur.

In addition to this, Meyer (2004, p.66) concludes that computerizing spoken and written texts for inclusion in a corpus is a very labor-intensive part of creating a corpus. As for genre, Biber (1988, cited in Vincent 2009, p.125) illustrates that *'genre refers to classes of texts that are determined on the basis of external criteria relating to author's or speaker's purpose',* whereas text type refers to *'classes of texts that are grouped on the basis of similarities in linguistic form, irrespective of their genre classifications.'* So, *text types and genres do not necessarily overlap*; they should be *distinguished.*

A Framework for Lexical Analysis (FLA)

According to Vincent's (2009) assumption,

> *The corpus-based lexical resources for lexical acquisition should facilitate the integration of the instantiations of the word and the company it keeps in the corpus with a structured lexical data/ knowledge base system, which records the various levels of linguistic information about the word. Accordingly, these various levels of information in a corpus informed lexicon can be obtained through the mediation of a bottom-up (corpus-processing) and a top-down (knowledge-based) approach.* (p.86)

Vincent (2009, p.86) points out the principles of the LFA as follows:

1-The lexicon is the central repository of linguistic knowledge, and so any analysis of language should take the word/lexeme as the central unit within which grammatical and lexical information are integrated.

2-This linguistic information is described using categories that may be regarded as common to varied linguistic theories.

3-Although linguistic and cognitive structures have their respective distinctiveness, they are not dissimilar, unless proven otherwise.

4-Linguistic knowledge may be viewed as declarative knowledge

5-A complementary view to language being structured as declarative knowledge is that the lexicogrammar may also be viewed as inherently probabilistic: statistical information such as frequency, mutual information and Z-scores for co-occurrence knowledge should therefore be recorded as facts in the lexicon

6-Lexicographic information can be derived from the observation of language in use.

7-Lexical knowledge is derived from analyzing how words are used. To derive this knowledge, the corpus used should be representative of the phenomenon under study.

8- It is useful to process a corpus first by using the appropriate corpus tool to economize human resources, the processing is to be done in an automatic manner if possible. However, for a small corpus, existing automatic methods of lexical acquisition which rely on statistical methods cannot always be applied because the statistics are unreliable for small bodies of data. So, human intervention is inevitable: human intervention also serves to ensure that the tools have been appropriately applied.

9-A lexical frame is used as the basic data structure for the lexical entry. The frame is structured through a combination of 'letting' the processed data suggest what these types of frames should be and specifying a sufficiently abstract level of analysis.

10- a lexical entry (and the lexicon) should be translated and organized into a lexical data/knowledge base system which has a computationally tractable and expressive format to act as a general, multifunctional resource for NLP applications.

RESEARCH METHODOLOGY

Based on the purpose of this study, exploratory mixed-method approach will be applied to develop a specialized word list for each major of oil industry training to support our trainees in their vocational training to be well prepared for their worksites in an oil company in Suadi Arabia . Further, the data collected in exploratory research is usually descriptive and helps me to identify patterns and trends, generate hypotheses, and develop a deeper understanding of the research problem. Accordingly, the qualitative results will be used to get a more in-depth understanding to the obtained quantitative findings. The quantitative and qualitative data are collected and analyzed in two sequential phases; the qualitative phase is built on the quantitative through purposeful sampling from the same population for conducting semi-structured interviews. Data collection process should start with completing the questionnaire for the quantitative method followed by the qualitative method as the quantative represents the major feature of my data collection process grounded in the purpose of the study. The mixing of the two methods occurred at two stages: first, while designing the interview protocols and choosing the participants for conducting follow-up interviews to do further exploration to the quantitative results, and second while integrating the final conclusions from both quantitative and qualitative phases at the interpretation and discussion stage of the study.

Data Collection

I will follow Meyer's (2004) concept to create the actual corpus once the basic outlines of the corpus are determined. This is a three-part process, involving the collection, computerization, and annotation of data. The first stage, collecting data involves recording interviews, gathering written texts or written course books, obtaining permission from speakers and writers to use their texts, and keeping careful records about the texts collected and the individuals from whom they were obtained. These collected data are computerized depending upon whether the data are spoken or written. Recordings of interviews with the technical instructors and recording some major-based classes, for example pipefitting, need to be manually transcribed using either a special video recorder that can automatically replay segments of a recording, or software that can do the equivalent with a sample of speech that has been converted into digital form. Written texts that are not available in electronic form can be computerized with an optical scanner and accompanying OCR (optical character recognition) software, or they can be retyped manually.

As mentioned in the preceding part, one purpose of this study is to develop a specialized technical word list for each major at Energy Tech college in Saudi Arabia. Accordingly, I am planning to use corpus-based data along with expert opinion on the technicality of words. The first steps are to analyze the written corpus and spoken recorded sessions applying frequency principles to identify items for the word list. Further, Hunston (2022, p.117) pinpoints the importance of quantitative measuring through applying the key word approach as she states *"Keywords are words that are significantly more frequent in the corpus being studied than in a more general corpus. Identifying keywords can give information about what is distinctive about one set of texts in comparison with another*." For this reason, I will compare the given specialized words with another ESP corpus-based by Saudi Aramco Company, the biggest oil industry company around the world.

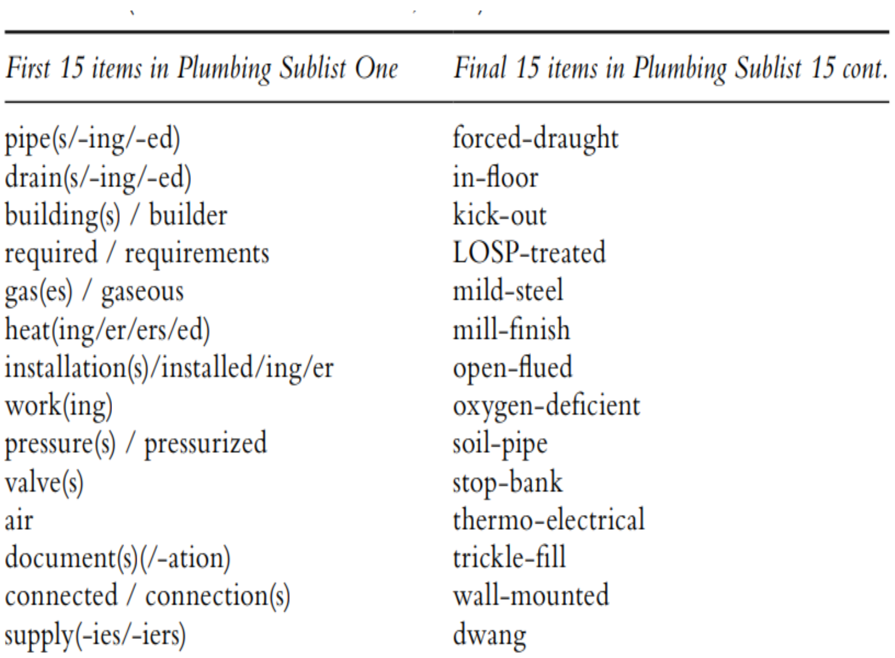
The next steps are to seek the expert opinion on the meaning of high-frequency items resulted in the initial analysis explained in the preceding paragraph. To do this, the technical instructors should be asked to rate the technicality of a sample of words through completing a 3-scale survey (2-0); they should award scores of 2 = the technical meaning in oil industry, score of 1= related to oil industry, but not very technical, and 0 score= not technical at all. After ranking all the possible items from the written corpus, maybe over 300 items, then the spoken corpus should be investigated for more specialized oil industry lexis. For the sake of further checking, dictionaries and glossaries for definition could be consulted where possible as well as the written and spoken corpora for evidence of how the words were used in context and technical meaning.

Data Analysis & Discussion

I am planning to replicate the study of Coxhead and Demecheleer's (2018, cited in Jablonkai and Csomay (2023) but using Oil Industry Word list in order to meet my target Saudi context. The first quantitative analysis should be used to measure the statistical aspects of the corpus including the keywords, frequency, and concordance. To illustrate this, I will use the examples of Plumbing Word List of Coxhead and Demecheleer's (2018, cited in Jablonkai and Csomay (2023):

> *Coxhead and Demecheleer's (2018) Plumbing Word List contains 1465 types which are arranged by frequency in 14 sublists of 100 items and one list of 65 items. [Table 1] shows the first 15 items in the most frequent sublist, beginning with pipe(s/-ing/-ed) and the last 15 items in the Plumbing Word List from the lowest frequency sublist, ending with wingback. Note the most frequent items in the column on the left include words which are commonly used in everyday English, e.g. air, building, and required, but also have a specialized meaning in plumbing. The items in the right column are not often encountered outside the field and clearly require specialized knowledge, e.g. LOSP-treated and dwang.* ***(*** p. 200***)***.

Figure 4: The 15 most and least frequent items in the Pluming Word List, Coxhead & Demecheleer, 2018, cited in -Jablonkai and Csomay, 2023, p. 200).



First 15 items in Plumbing Sublist One	*Final 15 items in Plumbing Sublist 15 cont.*
pipe(s/-ing/-ed)	forced-draught
drain(s/-ing/-ed)	in-floor
building(s) / builder	kick-out
required / requirements	LOSP-treated
gas(es) / gaseous	mild-steel
heat(ing/er/ers/ed)	mill-finish
installation(s)/installed/ing/er	open-flued
work(ing)	oxygen-deficient
pressure(s) / pressurized	soil-pipe
valve(s)	stop-bank
air	thermo-electrical
document(s)(/-ation)	trickle-fill
connected / connection(s)	wall-mounted
supply(-ies/-iers)	dwang

Jablonkai and Csomay (2023) note the differences in the amount of technical vocabulary in spoken and written texts in the plumbing study cited in Jablonkai and Csomay (2023). This is very clear in the extracted text in Table 2. It contains a short section of an interaction in class from the spoken plumbing corpus between a plumbing tutor and students in which they are discussing foul water and the Building Act of New Zealand.

Figure 5: A Plumbing tutor and students talk about foul water – Plumbing Word List adapted from Jablonkai

<Tutor:> (Turn 1) ***G13 Foul*** *water. What does* ***foul*** *water cover? What do you think it covers?*

<Student > (Turn 2) ***Sewage****.*

<Tutor:> (Turn 3) ***Sewage*** *and?*

<Student:> (Turn 4) ***Grey*** *water.*

<Student:> (Turn 5) ***Untreated****.*

<Tutor:> (Turn 6) ***Grey*** *water or? we don't actually use the term* ***grey*** *water yet as such, so what is it? What do we call it? We are referring to the Australian* ***standards strictly. Grey*** *and black water. So what do we?, especially when we are referring to* ***the building code****, what's the* ***grey*** *water referred to?*

< Student:> (Turn 7) ***Untreated****?*

<Tutor:> (Turn 8) No. Ok we've got ***soil fixtures****, haven't we? And that would be black water, so that's your* ***toilets*** *and* ***urinals****, and* ***grey*** *water would be?*

<Tutor:> (Turn 9) Oh your kitchen.

< Student:> (Turn 10) from the ***sink****.*

<Tutor:> (Turn 11) Yup. Shower, ***bath****, so what do we call it? No? The sanitary fixtures?*

< Student:> (Turn 12) ***Waste*** *water*

<Tutor:> (Turn 13) ***Waste*** *water, exactly. That's called?... we call that waste water. Ok? Now, as always when you are reading the* ***G13****, you'll be trying to get information, where do we start?*

<Student:> (Turn 14) at the start?

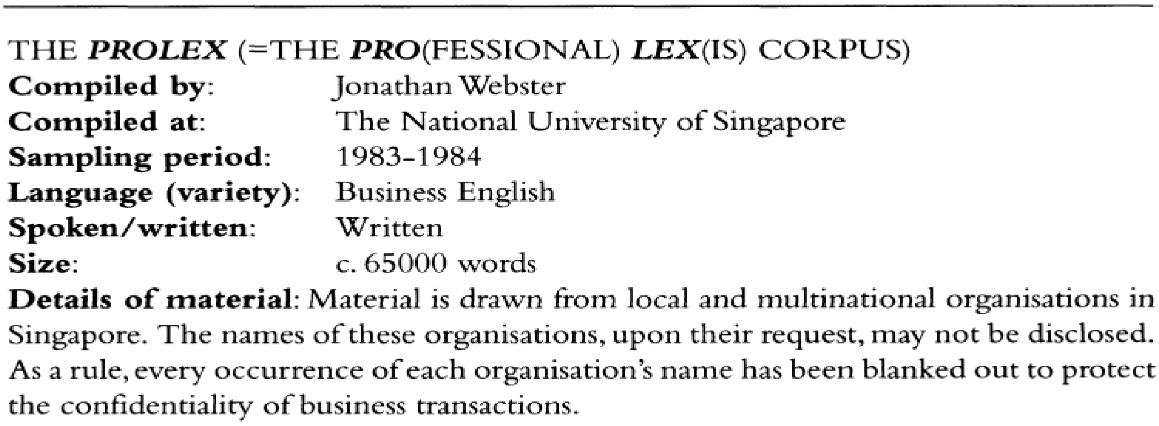
and Csomay (2023, p. 200)

As shown in figure 5 above, Items in the Plumbing Word List are highlighted in bold. There is roughly one word from the list per line, covering just over 16% of the words in the text. This coverage is higher than the whole plumbing spoken corpus reported by Coxhead and Demecheleer (2018) at 11.59%. Also, the Plumbing Word List items are often repeated, *sewage* in Turns 2 and 3, and *grey* as in *grey water* in Turns 4 and 6.

The Prolex Corpus = The Pro (fessional) Lex (is) Corpus

The first sublanguage corpus, the PRO (fessional) LEX (is), of business English texts, was gathered between 1983 and 1984. Samples of business English correspondence from both local and multinational firms in Singapore were collected with the result that the PROLEX corpus amounts to a text base of 566 samples which total approximately 65,000 words. More details are included in table 3 below.

Figure 6: The PROLEX corpus: a summary (Vincent, 2009, P.126)



THE ***PROLEX*** (=THE ***PRO***(FESSIONAL) ***LEX***(IS) CORPUS)

Compiled by: Jonathan Webster
Compiled at: The National University of Singapore
Sampling period: 1983–1984
Language (variety): Business English
Spoken/written: Written
Size: c. 65000 words
Details of material: Material is drawn from local and multinational organisations in Singapore. The names of these organisations, upon their request, may not be disclosed. As a rule, every occurrence of each organisation's name has been blanked out to protect the confidentiality of business transactions.

the objectives of the PROLEX project,

As outlined by Webster (1984, 1986, cited in Vincent 2009:126), some objectives of the PROLEX project are as follows:

> *1. to provide a syntactically sophisticated computer-processed frequency-count of professional lexis with systematic indications given of the social and geographical source of the form.*
>
> *2. to classify text samples according to those typological categories deemed to be sociolinguistically significant.*

3. to discover the collocational range of those lexemes constituting the professional lexis under review,

4. to develop PROLEX into an expert system for word knowledge, not just another on-line dictionary. (p.126)

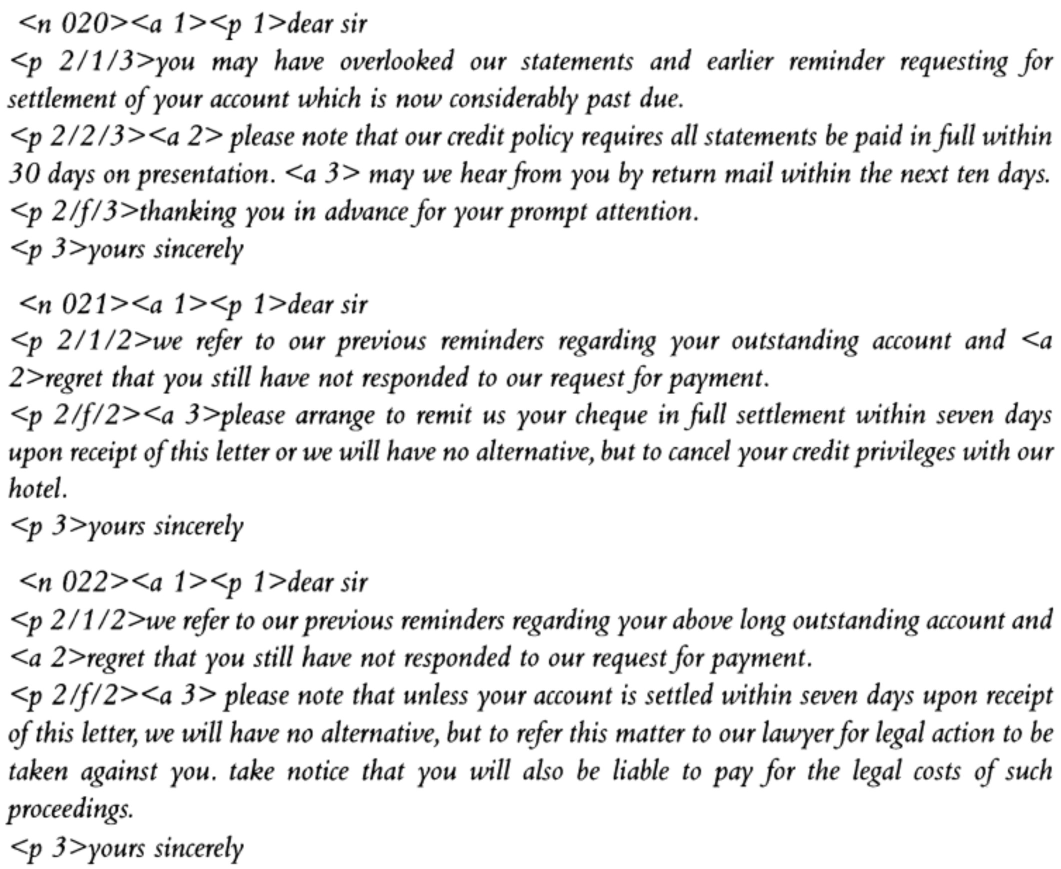
The PROLEX Project Principles

Webster (1984, 1986, cited in Vincent 2009, P.127) highlighted the following principles for gathering and analyzing a sublanguage corpus:

- Firstly, defining the system of rules underlying the sociolinguistic competence of members of this socially defined group.
- Secondly, indicating the human sociopsychological reality by organizing the representation of professional lexis by frequency, collocational range, and semantic nesting.
- Thirdly, providing a syntactically sophisticated, computer-processed frequency-count of professional lexis with systematic indications given of the social and geographical source of the form, (pragmatic tagging). The annotation of a corpus is needed to prepare it to be useful for further linguistic purposes.
- Fourthly, classifying text samples according to those typological categories determined to be sociolinguistically significant'. Pragmatic knowledge is an essential part of the lexicon's task of relating the word to its use in context.
- Fifthly, the corpus should be processed for its lexical content and then 'transformed' into a lexicon that can eventually be made available to various professionals and learners by virtue of its reusability.
- Sixthly, using the corpus as a lexical resource means a need to store and organize such a corpus-derived lexicon into a lexical data/knowledge base.

Figure 7 below shows some sample texts extracted from the PROLEX corpus which has been pragmatically tagged (an explanation of the pragmatic codes contained within angled brackets)

Figure 7: Sample texts from the PROLEX corpus, adapted from Vincent, 2009, P.128)



<n 020><a 1><p 1>dear sir
<p 2/1/3>you may have overlooked our statements and earlier reminder requesting for settlement of your account which is now considerably past due.
<p 2/2/3><a 2> please note that our credit policy requires all statements be paid in full within 30 days on presentation. <a 3> may we hear from you by return mail within the next ten days.
<p 2/f/3>thanking you in advance for your prompt attention.
<p 3>yours sincerely

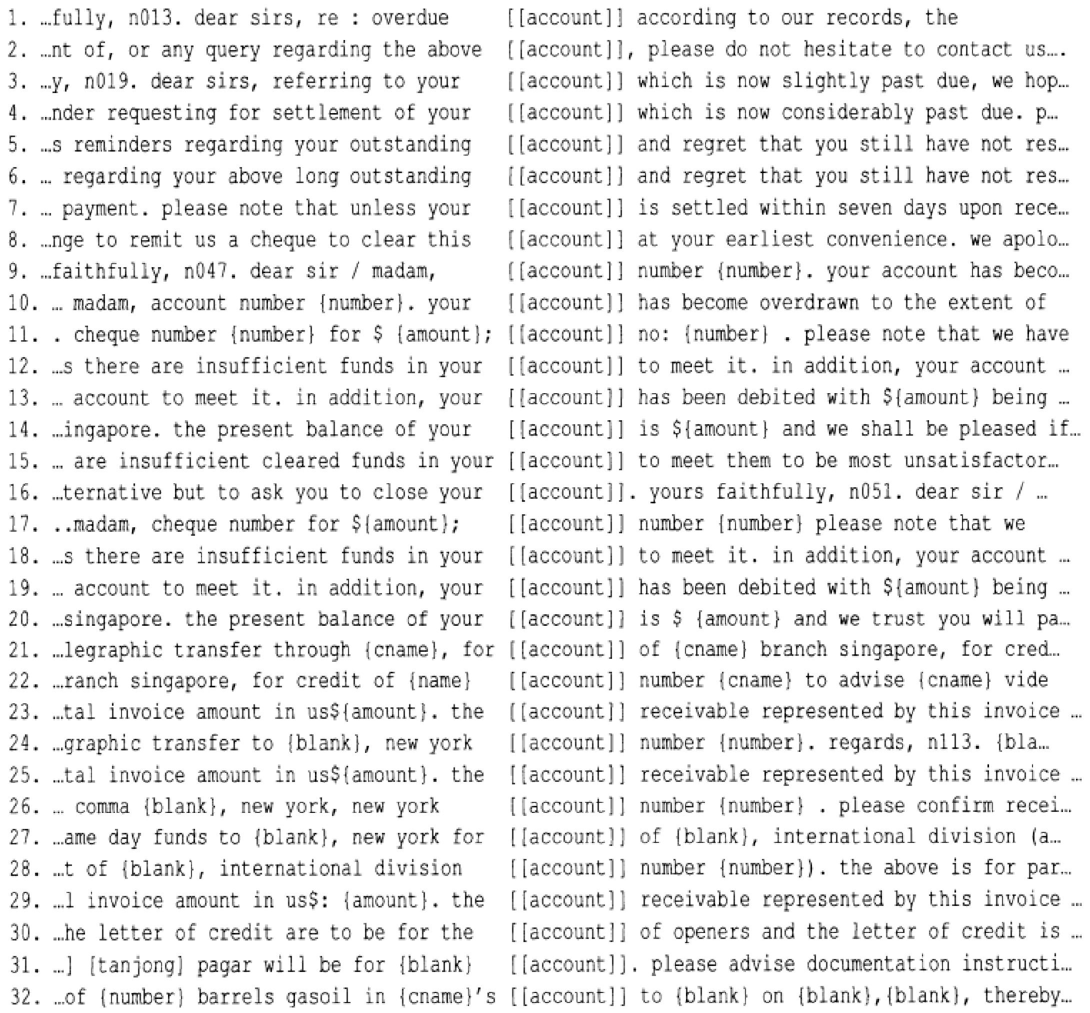
<n 021><a 1><p 1>dear sir
<p 2/1/2>we refer to our previous reminders regarding your outstanding account and <a 2>regret that you still have not responded to our request for payment.
<p 2/f/2><a 3>please arrange to remit us your cheque in full settlement within seven days upon receipt of this letter or we will have no alternative, but to cancel your credit privileges with our hotel.
<p 3>yours sincerely

<n 022><a 1><p 1>dear sir
<p 2/1/2>we refer to our previous reminders regarding your above long outstanding account and <a 2>regret that you still have not responded to our request for payment.
<p 2/f/2><a 3> please note that unless your account is settled within seven days upon receipt of this letter, we will have no alternative, but to refer this matter to our lawyer for legal action to be taken against you. take notice that you will also be liable to pay for the legal costs of such proceedings.
<p 3>yours sincerely

The PROLEX Lexicon

As shown by Vincent (2009), it can be noted that the progression of a harsher, more insistent demand for settlement of the account concerned in the sample text (Fig 4) above. The series of letters first begins with reference to the *'overdue*' account and moving progressively *to 'slightly past due'* account, 'considerably past due' account, *'outstanding*' account, and finally *'long outstanding'* account. However, a concordance listing is necessary to show systematically the use of the lexical item: the following concordance listing obtains for the lexeme account.

Figure 8: Concordance listing of account, from the PROLEX corpus, Vincent 2009, p.130)



```
1. …fully, n013. dear sirs, re : overdue     [[account]] according to our records, the
2. …nt of, or any query regarding the above  [[account]], please do not hesitate to contact us….
3. …y, n019. dear sirs, referring to your    [[account]] which is now slightly past due, we hop…
4. …nder requesting for settlement of your   [[account]] which is now considerably past due. p…
5. …s reminders regarding your outstanding   [[account]] and regret that you still have not res…
6. … regarding your above long outstanding   [[account]] and regret that you still have not res…
7. … payment. please note that unless your   [[account]] is settled within seven days upon rece…
8. …nge to remit us a cheque to clear this   [[account]] at your earliest convenience. we apolo…
9. …faithfully, n047. dear sir / madam,      [[account]] number {number}. your account has beco…
10. … madam, account number {number}. your   [[account]] has become overdrawn to the extent of
11. . cheque number {number} for $ {amount}; [[account]] no: {number} . please note that we have
12. …s there are insufficient funds in your  [[account]] to meet it. in addition, your account …
13. … account to meet it. in addition, your  [[account]] has been debited with ${amount} being …
14. …ingapore. the present balance of your   [[account]] is ${amount} and we shall be pleased if…
15. … are insufficient cleared funds in your [[account]] to meet them to be most unsatisfactor…
16. …ternative but to ask you to close your  [[account]]. yours faithfully, n051. dear sir / …
17. ..madam, cheque number for ${amount};    [[account]] number {number} please note that we
18. …s there are insufficient funds in your  [[account]] to meet it. in addition, your account …
19. … account to meet it. in addition, your  [[account]] has been debited with ${amount} being …
20. …singapore. the present balance of your  [[account]] is $ {amount} and we trust you will pa…
21. …legraphic transfer through {cname}, for [[account]] of {cname} branch singapore, for cred…
22. …ranch singapore, for credit of {name}   [[account]] number {cname} to advise {cname} vide
23. …tal invoice amount in us${amount}. the  [[account]] receivable represented by this invoice …
24. …graphic transfer to {blank}, new york   [[account]] number {number}. regards, n113. {bla…
25. …tal invoice amount in us${amount}. the  [[account]] receivable represented by this invoice …
26. … comma {blank}, new york, new york      [[account]] number {number} . please confirm recei…
27. …ame day funds to {blank}, new york for  [[account]] of {blank}, international division (a…
28. …t of {blank}, international division    [[account]] number {number}). the above is for par…
29. …l invoice amount in us$: {amount}. the  [[account]] receivable represented by this invoice …
30. …he letter of credit are to be for the   [[account]] of openers and the letter of credit is …
31. …] [tanjong] pagar will be for {blank}   [[account]]. please advise documentation instructi…
32. …of {number} barrels gasoil in {cname}'s [[account]] to {blank} on {blank},{blank}, thereby…
```

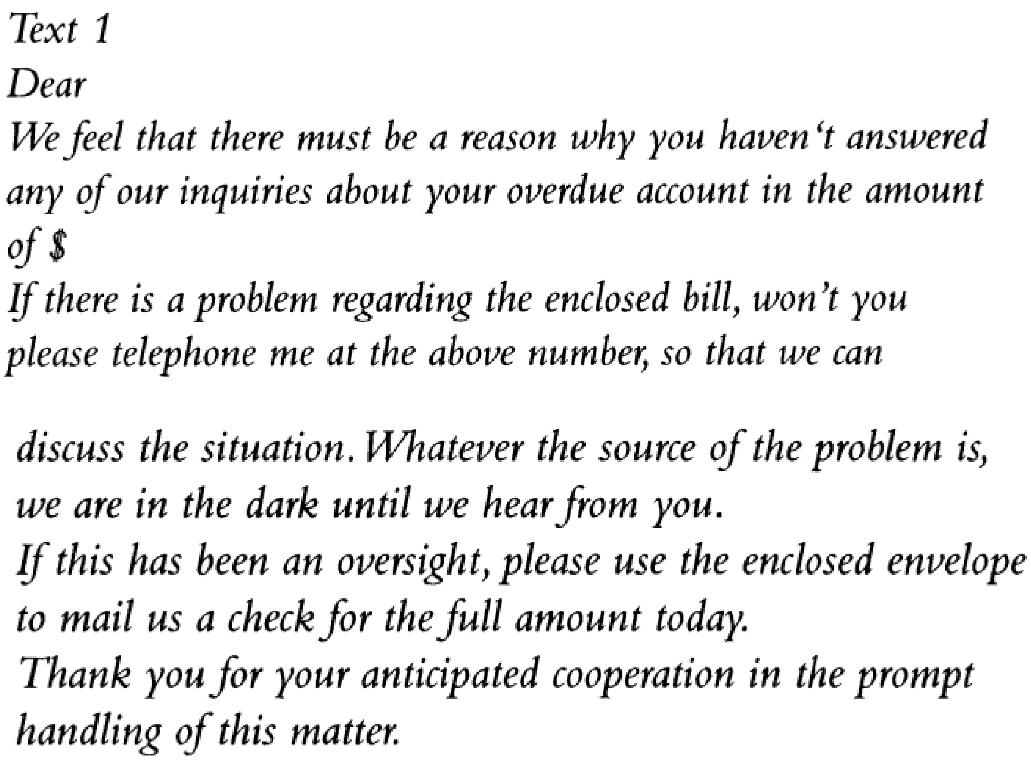
This concordance listing above shows several codes used in the PRO LEX corpus. These codes have been used as a condition for obtaining sensitive material from certain organizations; they requested that they should not be identified. For this reason, the abbreviation {cname }, enclosed in curly brackets, is used as a substitute for names of companies/institutions whose names may not be disclosed for reasons of confidentiality. In turn, { cname_obj} is used to denote the name of the company product concerned. {Blank} is used to delete sensitive person's names or account numbers. {Lname} is a composite term for a person's last name to maintain authorial anonymity. Similarly, ${amt} is used as a substitute for the actual sum of money transacted. There are also (self-explanatory) terms such as {date} and {city} in the corpus.

The Procompare Corpus

The PROLEX corpus would be compared with another similar corpus, The PROCOMPARE, for the sake of checking any inadequacy that might arise in its linguistic coverage of the domain of Business English. The

PROCOMPARE corpus was used as a control, providing a comparison with the PROLEX corpus and it is found in a supplement to PC-Shareware Magazine including Over 600 sample business letters and legal forms for your boiler plating, especially Accounting, Business, Legal, Employee, Product order, Sales letters, and some common forms. The various files have been merged into a large file, totaling approximately 60 thousand words, which is like the PROLEX corpus. Figure 6 below contains sample texts extracted for account from the PROCOMPARE corpus.

Figure 9: Sample texts from the PROCOMPARE corpus (Vincent 2009, P. 132)



Text 1

Dear

We feel that there must be a reason why you haven‘t answered any of our inquiries about your overdue account in the amount of $

If there is a problem regarding the enclosed bill, won't you please telephone me at the above number, so that we can discuss the situation. Whatever the source of the problem is, we are in the dark until we hear from you.

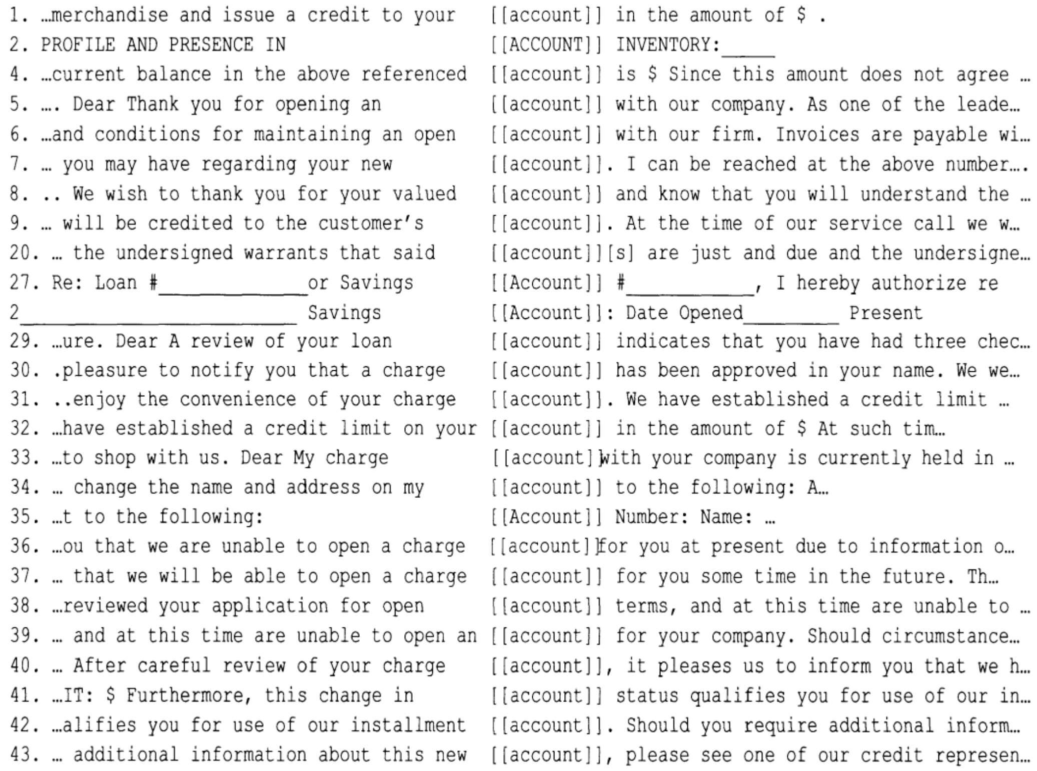
If this has been an oversight, please use the enclosed envelope to mail us a check for the full amount today.

Thank you for your anticipated cooperation in the prompt handling of this matter.

As concluded by Vincent (2009), the PROCOMPARE texts achieve the feeling of relative informality through, for example, the use of contractions (e.g. Text 1, haven't and won't) Also, the PROCOMPARE texts seem more formulaic (e.g. you, the second person pronoun, is blankly used, without referring specifically to any addressee). On the other hand, the PROLEX texts are more specific in intent since both addresser and addressee are known. Also, the PROLEX texts are more specific in content, and so perhaps give a clearer indication of what the business profession is writing about.

The PROCOMPARE lexicon

Figure 10: Concordance listing of account, from the PROCOMPARE corpus, Vincent (2009, P. 134)



```
1. …merchandise and issue a credit to your  [[account]] in the amount of $ .
2. PROFILE AND PRESENCE IN                  [[ACCOUNT]] INVENTORY:_____
4. …current balance in the above referenced [[account]] is $ Since this amount does not agree …
5. …. Dear Thank you for opening an         [[account]] with our company. As one of the leade…
6. …and conditions for maintaining an open  [[account]] with our firm. Invoices are payable wi…
7. … you may have regarding your new        [[account]]. I can be reached at the above number….
8. .. We wish to thank you for your valued  [[account]] and know that you will understand the …
9. … will be credited to the customer's     [[account]]. At the time of our service call we w…
20. … the undersigned warrants that said    [[account]][s] are just and due and the undersigne…
27. Re: Loan #_____________or Savings       [[Account]] #___________, I hereby authorize re
2_________________________ Savings           [[Account]]: Date Opened________ Present
29. …ure. Dear A review of your loan        [[account]] indicates that you have had three chec…
30. .pleasure to notify you that a charge   [[account]] has been approved in your name. We we…
31. ..enjoy the convenience of your charge  [[account]]. We have established a credit limit …
32. …have established a credit limit on your [[account]] in the amount of $ At such tim…
33. …to shop with us. Dear My charge        [[account]]with your company is currently held in …
34. … change the name and address on my     [[account]] to the following: A…
35. …t to the following:                    [[Account]] Number: Name: …
36. …ou that we are unable to open a charge [[account]]for you at present due to information o…
37. … that we will be able to open a charge [[account]] for you some time in the future. Th…
38. …reviewed your application for open     [[account]] terms, and at this time are unable to …
39. … and at this time are unable to open an [[account]] for your company. Should circumstance…
40. … After careful review of your charge   [[account]], it pleases us to inform you that we h…
41. …IT: $ Furthermore, this change in      [[account]] status qualifies you for use of our in…
42. …alifies you for use of our installment [[account]]. Should you require additional inform…
43. … additional information about this new [[account]], please see one of our credit represen…
```

It can be noted that the lexeme, account, must be customized in order to suit one's needs: for instance, there is a general reference to a 'said' account, which does not occur in the PRO LEX corpus; there are also more blank spaces to be filled in by the addresser of the template text.

Corpus Tagging

The following parts will elaborate corpus tagging occurred in the two English business corpora. As shown by Vincent (2009) that the CLAWS2 program was used for tagging the PRO LEX and PROCOMPARE corpora. This annotation involves 'assigning to each word in a text an unambiguous indication of the grammatical class to which this word belongs in this context.' CLAWS was able to perform the analysis successfully in a high degree of accuracy in the system, whereas some mistakes are made, and so human post-editing is needed. Here are some samples of tagged texts in figure 11 below.

Figure 11: Sample CLAWS output (uncorrected) for PROLEX texts on the lexeme account, Vincent (2009, p.137)

Vincent (2009) illustrates that the tagged texts were tagged using the following:

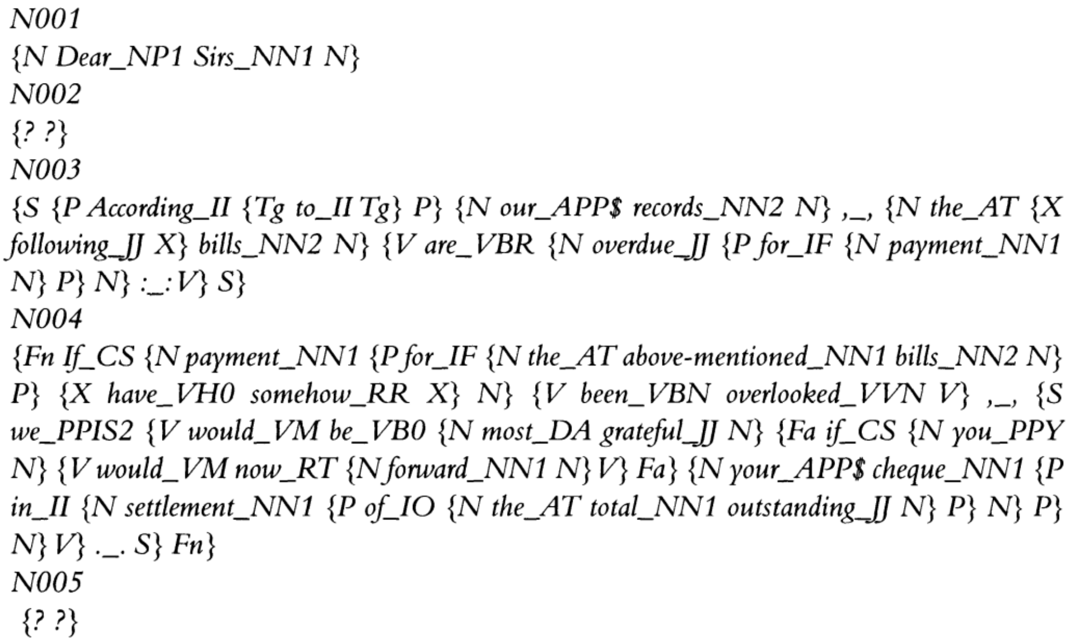
APP$ = possessive pronoun, pre-nominal; AT=article; CC=coordinating conjunction; CSA=as as conjunction; JJ=general adjecverb. (MC=cardinal number; NN=common noun; NNl =singular

> *common noun; NN2=plural common noun; NNJ=organization noun; NNT2=plural temporal noun; NNU=unit of measurement; PPIS2=we; RR =general adverb; VBO=base form 'be'; VBDZ=was; VBN=been; VBR =are; VVD=past tense of lexical verb; VVG=-ing participle of lexical verb; VVN=past participle of lexical verb.* (p.137)

Corpus (Syntactic) Parsing

Another method of processing the corpora involves the syntactic analysis (or parsing) of a corpus: from a parsed corpus, it is possible to retrieve information about more abstract grammatical categories which cannot be specified in terms of words or word-classes, for example, types of phrases or clauses. With parsing, one could automatically extract, for instance, valency information which would not be possible with just tagging. As concluded by Vincent (2009), using the same texts included in the above figures gives the output from the parser. It also contains the form of the texts required by the parser: sentence numbers are inserted before each sentence, the texts converted to mixed upper-lower case, and unwanted lines removed as shown in figure 9vbelow.

Figure 12: Sample parser output for PRO LEX texts on the lexeme account, Vincent (2009, p.140)



```
N001
{N Dear_NP1 Sirs_NN1 N}
N002
{? ?}
N003
{S {P According_II {Tg to_II Tg} P} {N our_APP$ records_NN2 N} ,_, {N the_AT {X
following_JJ X} bills_NN2 N} {V are_VBR {N overdue_JJ {P for_IF {N payment_NN1
N} P} N} :_: V} S}
N004
{Fn If_CS {N payment_NN1 {P for_IF {N the_AT above-mentioned_NN1 bills_NN2 N}
P} {X have_VH0 somehow_RR X} N} {V been_VBN overlooked_VVN V} ,_, {S
we_PPIS2 {V would_VM be_VB0 {N most_DA grateful_JJ N} {Fa if_CS {N you_PPY
N} {V would_VM now_RT {N forward_NN1 N} V} Fa} {N your_APP$ cheque_NN1 {P
in_II {N settlement_NN1 {P of_IO {N the_AT total_NN1 outstanding_JJ N} P} N} P}
N} V} ._. S} Fn}
N005
 {? ?}
```

As stated by Vincent (2009),

> *the ID/LP Parser uses the following non-terminal categories treebank (see Leech and Garside 1991; Black and Leech 1993): Fa (adverbial clause), Fe (comparative clause), Fn (noun clause), Fr (relative clause), G (genitive), J (adjective phrase), N (noun phrase), Nr (temporal adverbial noun phrase), Nv*

> *(non-temporal adverbial noun phrase), P (prepositional phrase), S (sentence), Tg (-ing clause), Ti (to-infinitive clause), Tn (past participle clause , V (verb phrase), ?null (unlabelled constituent). In addition, the symbols & and + respectively represent initial and non-initial conjuncts of a coordinate construction.* (p.141)

Vincent (2009) concludes that the Sharman parser is a step in the right direction for the analysis of texts: it is applicable to unrestricted text, employs a probabilistic grammar adapted to corpus analysis, and uses a parsing scheme which is as theoretically 'neutral' as possible, and is hence suited for a wider variety of uses.

Corpus Word-Extraction

For the sake of sorting the output from the CLAWS word-tagging, Vincent (2009, P.144) used Tony McEnery's Extraction Program which is designed to extract the frequency count of the possible tag(s) associated with each wordform in the tagged file. Therefore, a combination of the CLAWS program and this extraction program acts as a kind of lemmatiser, in the sense that the morphological variants of the word are indicated. This facilitates the structuring of the morphological parameter in the lexical entry.

Corpus Collocation

TACT (Text Analysis Computing Tools) assists in textual analysis by retrieving parts of text according to the wordform specified: it allows such facilities as concordance, frequencies, indexing and displaying the results in graphs, lists, and tables including a Z-score facility for the collation of word types related to the headwork. In statistics, the Z-score is a measure based on the standard deviation, involving the process of standardizing to facilitate the comparison of scores. As for the term "*collocational significance in a lexicographic sense refers to linguistically interesting pairs, which may or may not refer to either confidence intervals or expectations"* as stated by Vincent (2009, P.144).

the Z-score is applied to measure collocational strength as explained by Bradley (1990, cited in Vincent 2009, p.144) that "'*it takes the observed frequency of a word in the 'mini-text' ... and compares this with a theoretical frequency ... of occurrence within the same mini-text'.* Finally, Woods et al. (1986) define the Z-score as the standardized X-score described by the single formula Z = X1- X2 /s where s=standard deviation.

TACT uses the following formula, as suggested by Barron Brainerd (Bradley, 1990) Z= (Observed frequency of collocate -E) /s where P= frequency of collocate in full text / length of text, E= P x length of the mini-text. To conclude, Bradley, (1990, cited in Vincent 2009) confirms that *'a higher Z-score means more significance of the co-occurrence in a statistical sense'.*

Corpus Semantic Tagging and Parsing

Vincent (2009) asserts that the processing of a corpus need not be restricted to grammatical analysis and semantic tagging (of a corpus) is a technique still in its infancy. Nevertheless, semantic analysis is clearly an important level to apply to the extraction of lexicographic information from corpora, and major advances are being made to strengthen this aspect.

Corpus Pragmatics

Vincent (2009) explains how to process pragmatic information (i.e. 'meaning in interaction' as follows:

> *The first step is to exclude such formulaic texts before we process the corpus in this manner. The second step is to detail pragmatic information by referencing, either manually by hand or using an automatic computer program, the remaining texts for such information. By first processing the corpus for pragmatic information, we can later derive a lexicon that will 'mirror' speaker and hearer's knowledge, thus working towards the achievement of an all-inclusive lexicon.* (p.145)

In addition, Vincent (2009) illustrates the seven parameters of adding pragmatic facts to the PROLEX corpus only as follows: the referencing of each sample text in the corpus has specified the type of correspondence, the business organization from which the sample has been obtained, the function or purpose of the text, the speech acts and discourse moves involved, and the specification of which part of the text in which a lemma is found. The seven categories may be named as follows:

- The first label, the S (Subject) label, might perhaps be better called the F (Functions) label since it contains different functional categories.

- the second label, the C (Company) label, are listed the names of the business organizations

- The third label, the X (exchange) label, refers to the discoursal moves of either: 1. initiate or 2. Respond.

- speech acts under the fourth label, the A (Acts) label: acknowledge, inform, enquire/request/direct.

- The fifth label, the T (Type of Text) label, is used to Specify whether the word-form occurs as part of a 1. letter 2. telex 3. Memorandum

- the P (Part of Document) label, divides the text into different sections,

- The seventh label, the N (Number) label,

- The final label, the E (Error) label, is used for typographical and linguistic errors found in the original texts.

The COCOA format for referencing the PROLEX corpus.

COCOA (word COunt and COncordance on Atlas) format is adaptable for referencing programs. As explained in the preceding section, the typological categories are used to reference the corpus through using each number corresponding to the respective category. The texts are referenced by placing the typological labels between angled brackets. For instance, in the label <S 1 >, S (subject) indicates the category label, and number 1 indicates the value (i.e. Advertising) and so forth. Such formulae are inserted into the text where appropriate.

Structuring Lexical Entries from the PROLEX and PROCOMPARE Corpora

Once the PROLEX AND PROCOMPARE corpora are processed for the various types of linguistic information, it becomes much easier to structure a lexical entry from them. As stated by Vincent 2009. 149), "*not only can collocational, syntactic, semantic, and pragmatic information be derived, but frequency information detailing which of several related wordforms (*e.g. advise, advising, advised, advises) is most basic can also be obtained. The process of structuring a lexical entry should not obscure the relationship between the entry and other words. Accordingly, Mel'cuk and Zholkovsky (1988, cited in Vincent 2009, p.149) emphasized the notion of a vocable which they defined to be 'a *family of dictionary entries for lexemes which are sufficiently close in meaning and which share the identical stem*'.

Furthermore, Vincent (2009, p.150) developed four basic parameters of information for the selected PROLEX and PROCOMPARE lexemes: *morphology, syntax, semantics, and pragmatics* - of which the first three may be regarded as word knowledge and the fourth, world knowledge.

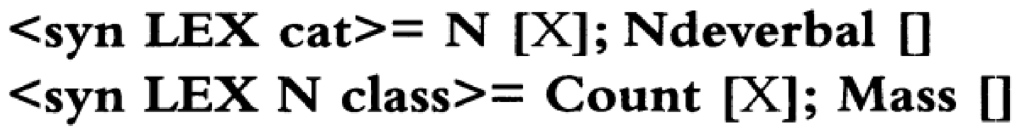
Morphology

Taking account (N) as the same illustrative example,

> Prolexnoun_LEX account (103) to be read as stated by Vincent (2009) as follows:
> *'Account is a PRO LEX noun lemma which consists of 103 occurrences of inflectional variants, i.e. of account (singular noun) and account (plural noun)'. In terms of morphology, the Morphology Frame for nouns thus specifies two basic slots:*
> *= "account (91) [NN1 91]"*
> *= "accounts (12) [NN2 12]"*
> *which indicates that the singular root/stem (F) form occurs 91 times, and its plural form (-s affix) 12 times. These frequency figures, with the respective CLAWS tag(s), are derived from McEnery's Word-Extraction Program which takes its input from a CLAWS output file.* (p.151)

Syntax

In terms of syntactic information, general information for the noun lemma account includes the following:



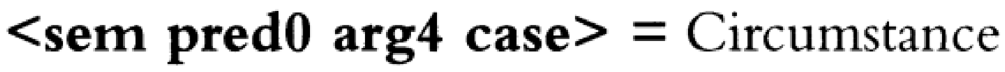
<syn LEX cat>= N [X]; Ndeverbal []
<syn LEX N class>= Count [X]; Mass []

This can be read as: account is a lexeme whose category is a noun (but not a deverbal noun- unlike claim) and whose class label is a count noun.

Vincent (2009) summarizes this process assuming that the main principle of ordering examples is to group them into the respective formal categories and then, within the formal category, sort these instances in terms of descending highest frequency. The frequency information shows that it collocates very strongly, for instance, with advice, this is also borne out by the Z-score ranking. The point here is that such types of information can be very useful for the lexicographer to extract for the creation of, say, a Dictionary of Oil Industry Collocations for the advanced learner.

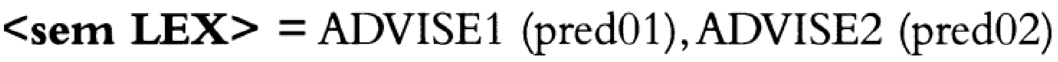
Semantics

This link between the semantic and syntactic (grammatical) parameters is therefore reflected, in the case of advise (v), in the following:



<sem pred0 arg4 case> = Circumstance

where advise is a predicate which takes four arguments, of which the Circumstance is classified as the fourth argument. Another example, these cases characterize the 'superordinate' predicate (predO) advise, from which these cases are inherited into the two senses of the verb in the PROLEX corpus.

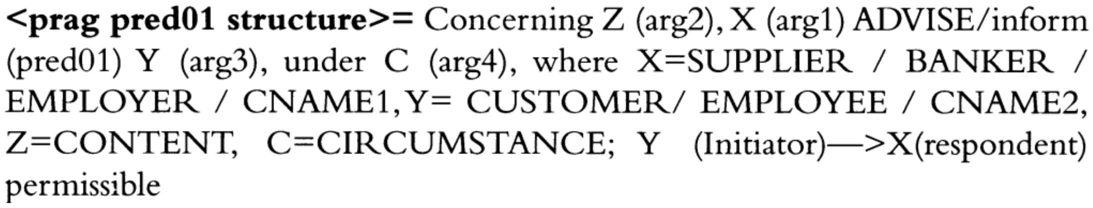


<sem LEX> = ADVISE1 (pred01), ADVISE2 (pred02)

ADVISE 1 has the sense of 'informing', while ADVISE2 takes the sense of 'directing /requesting' the addressee to a particular course of action.

7.4 Pragmatics

The contextual situation associated with the verb 'advise' is more widely reflected in a 'macro frame' / script, which takes a structure such as the following:



<prag pred01 structure>= Concerning Z (arg2), X (arg1) ADVISE/inform (pred01) Y (arg3), under C (arg4), where X=SUPPLIER / BANKER / EMPLOYER / CNAME1, Y= CUSTOMER/ EMPLOYEE / CNAME2, Z=CONTENT, C=CIRCUMSTANCE; Y (Initiator)—>X(respondent) permissible

Vincent (2009) proposed the following pragmatic interpretation of the verb 'advise' presented in the PROLEX corpus, it tends to be the institution/person initiating the discourse (be it the supplier, banker, employer, or more generally 'enamel') which/who does one of two things: either informing or directing the addressee (be it the customer, employee, or more generally 'cname2') regarding a particular state of affairs or course of action. Further, for advice, it is also permissible for the second party in the discourse, {cname2}, to initiate the piece of discourse as well. However, in the case of claim, which takes two related senses from the field of insurance.

STUDY FINDINGS

- Over 95% of all relevant lexical relations are obtained by examining collocates within a span of -5 and +5 not including punctuation marks.
- - Adopting a mix of statistics and linguistics/knowledge-based (hybrid) methodology seem promising for the acquisition of lexical information from English corpora.
- Various levels of information in a corpus informed lexicon can be obtained through the mediation of a bottom-up (corpus-processing) and a top-down (knowledge-based) approach.
- The PROLEX corpus texts are more specific in intent and content since both addresser and addressee are known, therefore they give a clearer indication of what the business profession is writing about. Meanwhile, the PROCOMPARE corpus texts seem more formulaic without referring to any addressee.
- CLAWS was able to perform the tagging analysis successfully, adding a grammatical class to the assigned word, in a high degree of accuracy in the system, whereas some mistakes are made, and so human post-editing is needed.
- A combination of the CLAWS program and Tony McEnery's Extraction Program acts as a kind of lemmatiser which facilitates the structuring of the morphological parameter in the lexical entry.
- The number of Specialized words in the written corpus could be more than spoken corpus.

CONCLUSION

The process of 'transforming' two corpora of business English into their respective lexicons which can be stored in a computational format. This process involves corpus lexical processing and acquiring the lexicons using a mix of a top-down and bottom-up approach. In terms of the top-down approach, it is necessary to use linguistically motivated categories for structuring the lexicon. As for the bottom-up approach, the (sublanguage) corpus has been suggested as an important complement to the dictionary (MRD) as a lexical resource. Such a lexicon, structured using the tenets of the LFA framework, is in a format amenable to computational storage in a lexical data or knowledge base which has the right tractability and expressiveness. Vincent (2009) concluded that this case study shows the process of 'transforming' two corpora of business English into their respective lexicons which can be stored in a computational format. This process involves corpus lexical processing and acquiring the lexicons using a mix of a top-down and bottom-up approach. In terms of the top-down approach, it is necessary to use linguistically motivated categories for structuring the lexicon; in terms of the bottom-up approach, the (sublanguage) corpus has been suggested as an important complement to the dictionary (MRD) as a lexical resource. Such a lexicon, structured using the tenets of the LFA framework, is in a format amenable to computational storage in a lexical data or knowledge base which has the right tractability and expressiveness. The results of this study could be utilized for conducting further investigation to expand the current limited areas of ESP and vocabulary research using corpora.

References

Hunston, S. (2022). *Corpora In Applied Linguistics*. New York: Cambridge University Press.

Jablonkai, R. & Csomay, E. (Eds) (2023). *The Routledge Handbook of Corpora and English Language Teaching and Learning*. New York: Routledge.

Krishnamurthy, R. (2008). Corpus-driven Lexicography. *International Journal of Lexicography*, Volume 21, Issue 3, September 2008, Pages 231–242. *http://dx.doi.org/10.1093/ijl/ecn028*

Lynne Flowerdew, L. (2012). *Corpora and Language Education*. Hong Kong: Palgrave Macmillan.

Meyer, C. (2004). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics :An Introduction*. Edinburgh: Edinburgh University Press

Vincent, B.Y. (2009). *Edinburgh Textbooks in Empirical Linguistics: Computer Corpus Lexicography.* Edinburgh: Edinburgh University Press

.