*Pietilä, P. & O-P. Salo (eds.) 1999. Multiple Languages – Multiple Perspectives. AFinLA Yearbook 1999. Publications de l'Association Finlandaise de Linguistique Appliquée 57. pp. 177–185.*

# USING QUANTITATIVE METHODS IN MEASURING STUDENTS' LEXICAL COMPETENCE

**Maarit Mutta**, University of Turku

This article discusses the use of quantitative vs. qualitative methods in a study which aims to measure the lexical competence of Finnish university students in French. It is a longitudinal study with an interlingual framework. The corpus consists of two parts; the first contains 106 essays, and the second 50, the latter representing the control group. The essays are analysed both qualitatively and quantitatively, the latter being the main concern. The aim of this quantitative approach is a more objective picture of the development of students' lexical richness. In order to make the required calculations, the unit to be used had to be determined, i.e. 'word'. Based on results of a pilot test made on one quantitative factor, 'word' is defined according to its orthographical definition. In summary, it is of primary importance to have the same criteria throughout the corpus and to remain faithful to your own criteria.

**Keywords**: French language, lexical richness, qualitative vs. quantitative method, word definition

## 1 INTRODUCTION

The main cognitive objective in language teaching today is the communicative competence containing linguistic (i.e. grammatical, lexical), pragmatic and discourse, sociocultural, and strategic competence (cf. Gaonac'h 1991; Faerch et al. 1984). Linguistic knowledge is based on four basic skills, namely the receptive skills, reading and listening, and the productive skills, speaking and writing. In my research (in preparation) I am interested in productive skills, particularly those of writing. However, these skills often overlap and therefore the limits between them are vague. According to Pontecorvo (1997: xviii) in a modern literate society it is almost impossible to separate the skill of writing from that of reading. In a sense, even if it were "theoretically possible to read without writing, it is almost

impossible to write without reading". As mentioned above, writing is the main object of interest here, although reading cannot be ignored as a part of the whole, especially as several studies prove that lexical acquisition of second language learning is closely related to reading (cf. Coady 1997).

In language learning and teaching, although it is important to know the grammatical rules of a language, it is equally, or even more essential to have good command of the appropriate vocabulary. Its role is strengthened by the fact that errors in vocabulary use are considered to be more disturbing for comprehension than grammatical errors (Carter 1987: 145). Adult second language learners especially may become frustrated if, due to lack of vocabulary, they are unable to express their ideas as freely and easily as they wish. This is one of the reasons why I wanted to examine the significance of vocabulary in second language acquisition in a particular situation, namely that of second language learning at university level. The study has an interlingual framework and its aim is to describe written competence, in particular the lexical competence of Finnish university students of French. The purpose is to discover how a stay in a French university environment (part of an ERASMUS exchange programme) influences the development of students' lexical richness and in what way; therefore it is also a longitudinal study. In summary, it is a question of verifying, for instance, the following hypotheses based on a preliminary study of the corpus:

1. A study period at a French university improves the students' lexical knowledge more effectively than a simple stay in France
2. Production becomes more natural, as, among other things, the use of discourse particles increases
3. Some oral features are filtered into written production

## 2 CORPUS

The corpus consists of two parts; the first contains 106 essays (2 x 53) by Finnish students of French written in an exam called **general language**

**know-ledge exam**[1]. Half have been written before a linguistic stay at the University of Aix-en-Provence (one of our exchange universities) between 1987 and 1993; the second half have been written after a stay which normally lasts one or two terms. These students participate in normal university teaching, primarily meant for native French students. The size of the corpus depends on the methods used, as a sufficiently large corpus is required to enable the drawing of reliable and definite conclusions of students' interlanguages and to show whether these phenomena are statistically significant. The second corpus contains 50 essays (2 x 25) of students of French who have not been to a French university, but have in fact spent at least one month in a French speaking country[2], this being included as a compulsory part of their second level studies. This latter group represents the control group and serves to verify the hypotheses on the first corpus.

## 3 QUANTITATIVE VS QUALITATIVE ANALYSES

Interlingual studies are very often of a qualitative nature. This work is also partly qualitative, but the main concern is a quantitative study. This type of approach seems to arouse a lot of controversial discussion, as is the case within the entire Arts field, as opposed to exact sciences (cf. Suomela-Salmi 1997: 16–17). Chomsky as well as Herdan (one of the main architects in lexical statistics) already had a heated discussion on this matter because of too narrow theoretical frames due to epistemological reasons (Ménard 1983: 2; cf. Dugast 1980: 7). However, many researchers in applied linguistics have had recourse to quantitative methods with successful results. For instance, Svartvik (1992: 9) states that

---

[1] This exam (*Kielitaitokontrolli*) aims to measure students' linguistic knowledge from the receptive and the productive point of view. It includes six different tests: the first two concentrate on oral language, the rest on knowledge of written French, which, e.g., consists of multiple choice questions on grammar and vocabulary, and of writing a composition on a given subject. The students must pass the exam twice during their studies; at the end of the first level (i.e. the end of the first year or during the first term of the second year), then at the end of their second level.

[2] The information has been gathered from students' reports on their linguistic stay. Most of these students have spent a month in a French speaking country, whereas a third have spent over six months working as an "au pair" and/or attending language courses for foreigners.

> Verifiability is a normal requirement in scientific research, and it is hard to see why linguistics (which is often claimed to be 'the scientific study of language') should be exempt from this standard mode of research procedure.

According to Ménard (1983: 2), statistics have always been a necessary means to linguists in many respects. Of course, neither Svartvik nor Ménard refer to interlingual studies, but this does not exclude the possibility of using applied statistics in the field in question. Cossette (1994: 1) follows the same line and claims that lexical richness, which is calculated with a statistical formula, is in fact an important concept in pedagogy. The advantage of using quantitative analyses in order to explain interlingual variation from the sociolinguistic point of view has also been shown (cf. Preston & Bayley 1996). With this quantitative approach, a more objective picture of the development of students' lexical richness in written production is searched for; as is generally known, a composition is a very subjective task to evaluate.

Lexical richness[3] contains here, besides a short view on students' errors, several quantitative factors, which together form a more complete picture of students' lexical competence. In order to examine students' vocabulary use in written production, I will have recourse to methods used by Linnarud (1986), but due to the different nature of my work I will apply them only partly (cf. Mutta 1995, 1997). Lexical factors calculated include:

1. Lexical variation (LV) – type / token ratio of lexical words
2. Lexical density (LD) – percentage of lexical words out of the total number of words
3. Lexical individuality (LI) – originality of lexis of each composition
4. Lexical sophistication (LS) – the level of difficulty of the vocabulary used

A computer program, **WordCruncher**[4]**,** will be used in the calculations. This quantitative analysis will be complemented by a qualitative one in order to give a global picture of the situation. This happens by analysing the vocabularies (of each student) in context and examining the interlingual development both individually and in the two corpora.

---

[3] Lexical richness is regarded here in a larger sense meaning the totality of students' vocabularies with their errors, etc.

[4] **WordCruncher** is a simple computer program that allows the user to retrieve and manipulate data from text files in a number of ways. For instance, one can study a word in its context and have such statistical reports as z-scores, the number of words and their frequency distribution.

# 4 DEFINITIONS OF ‘WORD’ – A PILOT TEST

In order to be able to make the required calculations, the unit used has to be determined. In an earlier study on English language (Mutta 1995) the term ‘word’ has already been discussed. It appears to be a clear concept to define, but, in fact, this is not the case. Ménard supposes that no linguist has managed to do this (1983: 20), at least not exhaustively. Does a word consist of only one word or group of words which together form the meaning of the word? What then is one word? A common definition is that it is a sound or combination of sounds that form a meaningful element, but this does not satisfy linguists, because ‘word’ is not in fact the smallest meaningful unit. Pergnier (1986: 13–15) claims that contemporary linguists agree on the character of this smallest meaningful unit that cannot be further divided, even if it is called by different names, e.g. *morpheme*, “monème” (A. Martinet). This is not the case with other linguistic units, including the term ‘word’.

One problem in defining the word is the diversity of approaches. We can have, among others, a linguistic, pedagogical or statistical approach, which all support different kinds of definitions for the unit to be treated. For instance, in linguistics, one can make a division between words that are graphically simple, but morphologically complex (e.g. derived or inflected forms) and units that are graphically complex, but form in fact a semantic unit (e.g. compound words, locutions or idioms) (Picoche 1992: 13–25). The latter are also called lexical units (cf. Bogaards 1994: 19). With graphically simple forms there arises, among other things, the question of the basic radical forms, whereas complex forms raise the question of criteria of limiting these lexical units by using tests of inseparability and substitution.

From the pedagogical point of view, it seems obvious that words should be learned as lexical units in larger contexts, and not as separate items, so that the different meanings and uses of a word would become clearer, or that the collocational uses of words would become more

adequate[5] (cf. Bogaards 1994; Coady 1997). When we speak of statistics, there is first to be found a difference between lexis and vocabulary[6], the former referring to *langue* (in a Saussurean sense) and the latter to discourse. The vocabulary is further divided into dictionary units, i.e. types (*vocable*) and occurrences, i.e. tokens (*mot*) (Cossette 1994: 4,102; Dugast 1980: 56). These occurrences can also be lemmatised so that different forms of a word are grouped under one lexical form (*lemme*). To a certain extent, all the approaches mentioned above are required for my study. In the following paragraphs, the choice for the definition of 'word' in the present study will be substantiated.

In consideration of what is said above, the word could be defined as a lexical unit, but this means that what should also be determined are words which form such a unit that they could be treated together, e.g. compound words and prepositions. But it could also be defined purely according to the orthographic definition: "a word is any sequence of letters which is bounded on either side by space" (Faerch et al 1984: 77). Carter (1987: 4) states that this "definition of a word is a practical common-sense definition" which "of course, refers to a medium of written language". This is useful to a certain extent when referring to English, but as Ménard puts it, there is a problem concerning the French orthography, which is all but coherent with its apostrophes and other pheno-mena typical of French (1983: 21).

Following some other studies of this kind, this orthographic definition was found to be useful for the purposes of an earlier study on written English (Mutta 1995; cf. also Elo 1993). This was also due to the computer program WordCruncher which calculated every separate item as one word. However, it was difficult to consider, for instance, compound words (e.g. *swimming pool*), complex prepositions (e.g. *because of*), conjunctions (e.g. *even though*) and pronouns (e.g. *each other*) as separate items, so they were counted as one word; this also included some other units such as proper nouns (e.g. names of countries) and fixed sentence adverbials (e.g. *of course*). In a way, a mixed definition of a word was thus used. Nevertheless, this was only partially satisfactory, because the problem of

---

[5] This means that, e.g., one can drink *strong tea*, but not * *heavy tea*, whereas a person can be *a heavy drinker*, but not * *a strong drinker* (Faerch et al. 1984: 95).

[6] These terms are used interchangeably for the greater part of the study.

limiting these lexical units arose during the work. In other words, where to draw the line between closely and more loosely related units? For instance, multi-word verbs (e.g. *put up with*) could have been counted as one unit, but their analysis with the WordCruncher program turned out to be too difficult, so they were calculated separately; however, they were treated manually in the qualitative analysis.

In the present study, I first intended to continue in the same way, but then thought I could solve the problem of limitation by using the pure orthographic definition of a word. In lexical research, one can also treat words in their graphic forms or as lemmatised forms (cf. Lavonius 1998). In order to discover whether the different definitions influence the results, a pilot test was made on one quantitative factor, i.e. lexical variation (LV = lexical type/token %), where ten compositions (2 x 5) representing the two different levels were chosen. The results are presented in Tables 1 and 2.

TABLE 1. Word according to earlier definition (mixed definition)

| student | LEVEL I | | | | LEVEL II | | | | % diff. |
|---|---|---|---|---|---|---|---|---|---|
| | number of words | type | token | LV % | number of words | type | token | LV % | |
| 1 | 165 | 72 | 99 | 72.7% | 187 | 75 | 98 | 76.5 | +3.8 |
| 2 | 175 | 81 | 96 | 84.4 % | 192 | 87 | 106 | 82.1 | -2.3 |
| 3 | 175 | 70 | 94 | 74.5 % | 181 | 84 | 105 | 80 | +5.5 |
| 4 | 196 | 97 | 107 | 90.7 % | 231 | 112 | 129 | 86.8 | -3.9 |
| 5 | 272 | 114 | 142 | 80.3 % | 315 | 137 | 173 | 79.2 | -1.1 |

TABLE 2. Word according to pure orthographic definition

| student | LEVEL I | | | | LEVEL II | | | | % diff. |
|---|---|---|---|---|---|---|---|---|---|
| | number of words | type | token | LV % | number of words | type | token | LV % | |
| 1 | 197 | 67 | 100 | 67 % | 205 | 74 | 102 | 72.6 | +5.6 |
| 2 | 197 | 85 | 104 | 81.7 % | 211 | 86 | 108 | 79.6 | -2.1 |
| 3 | 203 | 74 | 104 | 71.2 % | 203 | 86 | 115 | 74.8 | +3.6 |
| 4 | 220 | 95 | 114 | 83.3 % | 246 | 114 | 135 | 84.4 | +1.1 |
| 5 | 312 | 109 | 155 | 70.3 % | 357 | 136 | 191 | 71.2 | +0.9 |

In the first table, word is defined according to the mixed definition (cf. Mutta 1998), and in the second according to the pure orthographic definition, i.e., every unit surrounded by a space is a word; apostrophe forms were also separate units. A high LV implies that there is little repetition of lexical words, but this does not indicate the quality of the vocabulary, only that it is somewhat limited. Both tables indicate that those students having lower LV percentages at the first level (1 & 3) seem to profit more from their stay abroad because their LV improves at the second level. The other students seem to have percentages which are lower or only slightly higher at the second level as at the first, but this is probably due to the fact that in a longer composition it is more difficult to avoid repetition (e.g. number 5). In summary, one can conclude that the stay in France has had a more positive effect on students whose initial levels were lower. It can also be concluded from this pilot test that the results have similar tendencies in spite of the different definitions of **word**. It is of primary importance to have the same criteria throughout the corpus. According to Dugast (1980: 57), it is primarily important to be faithful to your own norm.

## 5 CONCLUSION

After all the discussion presented above, we can come to the conclusion that different definitions of the word according to linguistic/statistical, and pedagogical approaches can and must be used. It is important that, pedagogically speaking, the word is replaced by lexical units which are taught in meaningful contexts. But still, in order to perform the quantitative calculations it is more useful to have recourse to an orthographical definition. In the present study, a restricted number of units which cannot have their own separate significance (e.g. *parce que*) will be included in this group. Moreover, the forms will be left unlemmatised[7], except when dealing with lexical sophistication (i.e. the level of difficulty of the vocabulary used[8]). Lexical sophistication is partly manually treated and overlaps with the qualitative analysis. Furthermore, homographs are grouped under the same

[7] See for instance Lavonius (1998) for an orthographical definition in the larger sense and for lemmatised forms.

[8] It is quite evident that the level of difficulty of the vocabulary is a concept that can be discussed. Here it is attached to the frequency of words, as is often done.

word in calculations, but treated separately in the manually made qualitative analysis. Finally, the quantitative analysis is completed by a qualitative analysis which contains, in addition to that mentioned above, a short view on students' errors, a more detailed analysis of each student's vocabulary in context, and an examination of interlingual development both individually and in the two corpora.

I am aware of the fact that this type of approach can be criticized, and that according to other "norms" the results can be different. However, with this objective point of view on the development of lexical richness I hope to make my contribution to interlingual problematics.

## References


Bogaards, P.1994. *Le vocabulaire dans l'apprentissage des langues étrangères*. Paris: Hatier-CREDIF.

Carter, R. 1987. *Vocabulary. Applied linguistic perspectives*. London: Allen & Unwin Publishers Ltd.

Coady, J. 1997. L2 vocabulary acquisition through extensive reading. In J. Coady & T. Huckin (eds.) *Second language vocabulary acquisition*. Cambridge: CUP, 225–237.

Cossette, A. 1994. *La richesse lexicale et sa mesure*. Paris: Honoré Champion Editeur.

Dugast, D. 1980. *La statistique lexicale*. Genève: Editions Slatkine.

Elo, A. 1993. *Le français parlé par les étudiants finnophones et suédophones.* Turku: Turun yliopisto.

Faerch, C., K. Haastrup & R. Phillipson 1984. *Learner language and language learning*. Copenhagen: Gyldendalske Boghandel, Nordisk Forlag A.S.

Gaonac'h, D. 1987. *Théories d'apprentissage et acquisition d'une langue étrangère*. Paris: Hatier-CREDIF.

Lavonius, E. 1998. *Sanasto ranskan kielen ylioppilaskokeessa*. Turun yliopisto. Ranskan kielen ja kulttuurin pro gradu –työ.

Linnarud, M. 1986. *Lexis in composition.* Malmö: GWK Gleerup, Liber Förlag.

Ménard, N. 1983. *Mesure de la richesse lexicale*. Genève: Slatkine-Champion.

Mutta, M. 1995. *A study of the vocabulary used in entrance examination compositions*. Turun yliopisto. Englantilaisen filologian sivuainetutkielma.

Mutta, M. 1997. Lexical factors in entrance examination compositions. In H. Dufva (ed.) *FINLANCE* XVII. Jyväskylä: Centre for applied language studies, 80–91.

Mutta, M. 1998. La compétence lexicale des étudiants finnophones en français. In *Actes du premier Congrès des Romanistes Scandinaves pour étudiants en doctorat*. Lund : Lunds Universitets romanska institution, 11–18.

Pergnier, M. 1986. *Le mot*. Paris: Presses Universitaires de France.

Picoche, F. 1992. *Précis de lexicologie française.* First ed. 1977. Paris: Nathan.

Pontecorvo, C. 1997. Introduction: studying writing and writing acquisition today. A multidisciplinary view. In C. Pontecorvo (ed.) *Writing development: an interdisciplinary view*. Amsterdam: John Benjamins Publishing Company, xv– xxxi.

Preston, D. R. & R. Bayley 1996. Preface. In R. Bayley & D. R. Preston (eds.) *Second language acquisition and linguistic variation*. SiBil 10. Amsterdam: John Benjamins Publishing Company, xiii–xviii.

Suomela-Salmi, E. 1997. *Les syntagmes nominaux (SN) dans les discours économiques français: repères textuels*. Turku: Turun yliopisto.

Svartvik, J. 1992. Corpus linguistics comes of age. In J. Svartvik (ed.) *Directions in corpus linguistics*. Berlin: Mouton de Gruyter, 8–13.