LANCASTER UNIVERSITY

Department of Linguistics and
Modern English Language

MILITARY ENGLISH VERSUS GENERAL ENGLISH

A CASE STUDY OF AN ENGLISH PROFICIENCY TEST

IN THE ITALIAN MILITARY

Dissertation submitted in partial
fulfilment of the requirements
for the degree of

M.A. in Language Testing
(by distance)

Francesco Gratton
Supervisors: Prof. Charles Alderson/Richard West

July 2009

(18,017 words)

Abstract

In recent years, the use of corpora has proved to be a powerful tool in the field of language education. In addition, the field of testing has benefitted from the use of corpora as it allows, for instance, the development of academic vocabulary-size tests for non-native speakers of English entering tertiary education institutions.

This study has aimed at focusing on the use of corpora in English for Specialized Purposes (ESP) - in particular military English - and investigated whether the use of more job-related terminology in reading comprehension assessment has had a positive effect on the performance of test-takers, or whether it has brought on added difficulty.

The final results lead to believe that in a test with a high frequency of military-related terminology, the scores are negatively affected. The present research has also evidenced some limitations as regards the small sample of test-takers under scrutiny, the data-gathering method and the methodology adopted which should be more focused.

However, further research is needed to better understand how the use of specific terminology can truly and reliably reflect language ability in a military context.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBRIEVIATIONS

BILC	Bureau for International Language Coordination
DI	Discrimination index
ESP	English for Specialized Purposes
FV	Facility Value
IA	Item Analysis
JFLT	Joint Forces Language Test
TUI	Test Unificato Interforze (English Proficiency Test)
STANAG	Standardization Agreement
LSP	Language for Specialized Purposes
NATO	North Atlantic Treaty Organization
OSCE	Organization for Security and Cooperation in Europe
PfP	Partnership for Peace
SLP	Standardized Language Profile
SPSS	Statistical Package for Social Sciences
UN	United Nations

ACKNOWLEDGMENTS

Chapter 1

INTRODUCTION

This chapter will provide a rationale for the study being undertaken – i.e. a lexical research investigating whether specific terminology has incidence on a proficiency test administered to military personnel. Furthermore, it will provide an overview of how this study was conducted and how it is presented in this paper.

I first became interested in the topic of my dissertation when test-takers' initial feedback was returned following the administration of a new high stakes proficiency English test. The test had replaced the ten-year old proficiency test which had not only run its course but was also based on the former edition of the language proficiency scale STANAG 6001 in use in the military field.

As an officer in the Italian Army, I first encountered this test when appointed chief of the Testing Office of the Army Foreign Language School over 15 years ago. The school, besides offering language courses, is also the Official Language Certification agency of the Italian Armed Forces. Italian Military personnel are required to have a certified level of English from our school in order to qualify for deployment abroad and/or specific international positions within UN, NATO, OSCE, UE, etc. Since 1949 Italy has been committed to the North Atlantic Treaty Organization (NATO) along with 25 other Countries.

One of the main issues of the NATO multinational environment is the teaching and the assessment of languages. In 1966, the Bureau for International Language Coordination (BILC) was established by NATO members 'to disseminate to participating countries information on developments in the field of language learning' (Bureau for International Language Coordination, 2002). A major step that BILC undertook in 1976 was to lay down a set of language proficiency levels to be adopted by NATO, known as the Standardized

Agreement 6001 ed. 1, approved in 1976 (hereon referred to as STANAG 6001 Ed. 1). Under this agreement, all NATO Countries have committed to use these proficiency levels for the purpose of:

- meeting language requirements for international staff appointments;
- comparing national standards through a standardized table;
- recording and reporting, in international correspondence, measures of language proficiency if necessary by conversion from national standards. (North Atlantic Treaty Organization, 1976: 1).

Stanag 6001 Edition 1 prescribed six levels for the four skills of Listening, Speaking, Reading and Writing labelled as follows:

0   no practical proficiency

1   elementary

2   fair (limited working)

3   good (minimum professional)

4   very good (full professional)

5   excellent (native/bilingual)

After twenty years, NATO countries found that STANAG 6001 Ed. 1 descriptors were at times not detailed enough for an accurate and standardized assessment to be made across and within countries. Furthermore, the international geopolitical scenario had changed in the interim and the challenges military personnel faced while posted abroad had become multifaceted. Indeed, the world events following the fall of the Berlin Wall and the end of the Cold War have conferred new meaning and scope to military intervention in the so called theatre of operations and language learning and assessment have duly taken on new objectives to reflect these needs. The revision and subsequent drafting of the Stanag document is a further reflection of this new scenario.

As a result, in 1999 a BILC working group made of representatives from 11 NATO nations was assigned to revise the shortcomings of the first edition of the

STANAG and to develop what is known as the "Interpretation Document", finalized and approved in 2003. This document was appended to the original STANAG and is now known as "STANAG 6001 (Edition 2)" (Green & Wall 2005: 381).

This second edition provides testing teams with more detailed descriptors which prescribe the performance of language proficiency in terms of content, task and accuracy demands which aim not only to guide test developers and language tutors, but also to provide a common framework across NATO and PfP[1] countries with the description of standardized language proficiency performances.

In particular, edition 1 of the STANAG 6001 reading comprehension descriptors called for the successful candidate's skills at level three to be "adequate for standard test materials and most technical material in a known professional field, with moderate use of dictionary, adequate for most news items about social, political, economic, and military matters. Information is obtained from written material without translation" (STANAG ed 1 1976: A-3). The three planes of interpretation of this scale, i.e. content, task and accuracy are vague with performance standards only described in very scant details. The topical range from which texts can be extrapolated, however, were detailed enough for most test developers to have selected texts from most professional and technical fields limited of course to the knowledge of the test developers themselves.

Contrarily, edition 2 is definitely more detailed considering that the successful candidate at the same level three is described as having the reading comprehension skills to "…. Read with almost complete comprehension a variety of authentic written material on general and professional fields….demonstrates the ability to learn through reading…..comprehension is not dependant on subject matter…..contexts include news, informational and editorial items in major periodicals intended for educated native readers…."

---

[1] PfP: Partnership for Peace: is a NATO program launched in 1994 whose aim is to create trust between the same Alliance, non NATO European States and the Former Soviet Union. At present time there are 23 member states.

(STANAG Ed 2 2003: A-1-7). It is clear to see how test developers had not only more elaborate guidelines to follow when selecting texts from wider and more varied topic domains compared to the former edition, but they also had tasks that were clearly described, with which assessment of the successful candidate could be made. This also provided the more detailed accuracy demands specified for each level.

Every NATO country, including Italy, is required to develop its own national test to assess the language proficiency of its personnel; test results are reported using a four digit Standardized Language Profile (SLP). Each digit stands for the level the test taker has achieved in each linguistic skill and respectively in the order of Listening, Speaking, Reading and Writing. For example an SLP of 2-3-3-1 means 2 in Listening, 3 in Speaking, 3 in Reading and 1 in Writing. (Green & Wall 2005:380).

Upon approval of the second edition of the STANAG 6001 in 2003, the Italian Defence also felt it necessary to replace the ten-year-old 'Test Unificato Interforze', the English Proficiency Test (hereon referred to as TUI) which was based on the language descriptors of the first edition. The TUI was developed in 1997 to adhere to the STANAG proficiency requirements with an emphasis on specific military terminology to reflect what was described in levels 2 ad 3 as 'professional material' or 'job- related context' (STANAG 6001 ed 1. 1976).

On the contrary, the new edition of the STANAG and in particular, the amended "Interpretation Document", provided test developers with guidelines on which language functions typified the performance levels along with the topical domains and the tasks successful candidates at each level could carry out in accordance with the accuracy demands. Initial feedback from test takers at all levels stated that the new test was 'less military' in flavour.

The persistent comments on how test takers found the new test 'more difficult' because of its wider range of topical domains rather than a concentration on

military topics as in the former test led me to question if there was indeed a link between topic concentration and test performance.

In fact, while there have been extensive investigations into the impact of vocabulary knowledge, topic knowledge and test performance on a variety of tests, I could find little evidence of many studies undertaken in the military context, especially the Italian one.

This could be due to numerous reasons among which:
- the Italian Defence, despite its many years of experience in teaching and testing foreign languages, does not have a long history of research. The nature of a military establishment in Italy is such that it is administered by military personnel who may be often transferred to other positions. In addition, all civilians who teach and test foreign languages in military schools are on temporary contracts. Whilst this reality fosters a very mobile and dynamic working environment, it can also hinder the creation of fertile ground in which professional roots can grow, and upon which research activity would inevitably thrive.
- the uniqueness of the military context which is wary of sharing and divulging both test scores and/or material, considered and protected as 'military classified material' of the second level.[2]

Given this lack of research in the specific military environment I attempted to bridge this gap as the following paragraphs will detail.

1.1 Research context

Given the importance of the test scores and in order to provide more insight into evidence which could back up test takers' perceptions, I compared the reading comprehension components of the two proficiency tests: the T.U.I. and the Joint Forces Language Test (hereon referred to as JFLT). Although the two tests

---

[2] military material is classified according to its content and purpose. Foreign language test material is not the ranked at the highest level of classification; nevertheless, being valid for qualifications and career advancement, tests are considered very high stakes and their content carefully protected.

might seem very similar given that they are both proficiency tests consisting of 60, four-option multiple-choice items for the listening and reading comprehension components of the test, assessing all four skills at five levels of proficiency, under closer examination they are very different in content. The TUI presents a high frequency of job-related words (specifically military), included in the construct of the TUI specifications on the basis that this specific terminology was fundamental for the military personnel to have acquired in order to be considered qualified to work abroad: this was terminology they were likely encounter in theatre, given the military scenario at the time.

Nevertheless, taking into account that military personnel may come from different areas of competence or specializations such as the administrative, the medical, the engineering corps, to name but a few, consideration was given to the fact that the use of too specific military terminology might actually be biased towards certain candidates and hinder rather than facilitate the objectivity of test-takers' performances. As a result, when the test specifications for the JFLT were drafted, it was decided that using more 'neutral' vocabulary including matters of professional interest to all military personnel regardless of their professional background would be more appropriate, along with assessing the skills through the new fashionable 'geopolitical topics'. As a reminder, a new world scenario was in the making whereby Italian military personnel were, and still are, called upon to perform duties and tasks in collaboration with the local authorities of the country in which they are serving as well as other contingencies from all over the world – all using English as their working language. These duties and tasks entail the knowledge of not only specific military terminology but also and most importantly, of language functions to be able to deal with new authentic situations such as:

- carrying out patrol duties and delivering humanitarian aid to the local population using English to speak about the immediate environment on a concrete level as per the descriptors of STANAG ed. 2 level 2;
- collaborating with local authorities on the reconstruction of infrastructures (vital to war-torn countries) and on the training of the local army by using

English to speak about factual and abstract topics (as per the descriptors of STANAG ed. 2 levels two and three depending on the post assigned);

- negotiating, hypothesizing, supporting an opinion as per the descriptors of STANAG ed. 2 level 3 to be able to interact with the civilian population but also to perform diplomatic functions at the higher political and judicial levels.

On the basis of the above, the JFLT was developed with an emphasis on language functions, tasks and content domains typical of each level with less concentration on specific military terminology. Feedback collected during the trialling stages of the JFLT indicated that on the one hand, test takers who belonged to less specific military branches such as military physicians and veterinary doctors, were relieved that the new test contained fewer military topics and hence, military-specific vocabulary, whereas on the other hand, more operative test takers were surprised to find general topics e.g. daily news items, geopolitical issues with less emphasis on specific lexicon.

Although concurrent validity was established between the TUI and the new JFLT test takers' considerations encouraged me to find out if and to what degree the two tests actually differed in terms of the frequency of military topics and therefore of fewer specific terminology and if a difference did indeed exist, to what extent did it affect test takers' performance. To go about this, it was necessary to analyse in detail the exact terminology included in both tests. A careful study of not only the topics but especially the vocabulary in these topics needed to be conducted and triangulation with a more qualitative research method which could collect information from the test takers directly had to be conducted. Triangulation with the results of an interview aimed to probe into candidates' perceptions of the test as regards the topics and strategies they adopted whilst answering the items would either confirm that there was indeed a relationship or not between the incidence of military terminology and test scores.

This is of particular significance for me in my position as Head of the Testing Office given that the inferences that are made on the scores must reflect test takers' actual level of English proficiency. Most Officers and Non-Commissioned Officers qualify to work abroad as a result of their test scores and these scores should indicate if they are able to linguistically perform their duties in international environments. These include life-threatening situations for themselves and others, e.g. both military personnel and the civilian population.

I hope the results of my humble investigation can give some insight into whether, and to what extent, test scores are affected by the incidence of military terminology.

1.2. Overview

To answer my research question:

"Does less specific military terminology of the new Joint Forces Language test of the Italian Defence affect military personnel scores?",

I will proceed as follows: in chapter 2 I will provide a summary of what has been undertaken in the field of computational linguistics and especially corpora as this is pivotal in providing information on how terminology is categorized; I will then continue by providing a description of the detailed vocabulary thought to be strictly military in nature in relation to the two tests and then conduct an analysis. Chapter 3 will then illustrate how I administered the two tests at different times using the same group of selected test-takers. Test takers feedback received during an interview will be provided. I will analyze their scores by running descriptive statistics, classical item analysis and small-sample paired t-test to validate my hypotheses about differences existing between the two means.

In chapter 4, I will discuss the results and the statistical interpretations whereas Chapter 5 presents the conclusions, the limitations and implications for future research.

Chapter 2

LITERATURE REVIEW

The literature review in this chapter is presented in two sections.

The first section reviews literature related to the development of corpus, beginning with its first appearance in the field of linguistics and its development into a potentially useful tool in the field of language testing. In particular, some technical aspects of how a corpus is created, corpus typologies and coding procedures are described.

The second section of this chapter introduces literature related to issues of testing Military English in general and vocabulary along with readers' prior knowledge in topics, both of which have a bearing on my research.

Furthermore, a brief description of the language proficiency scale (Stanag 6001 ed. 1 and 2) follows. Finally, a discussion of the issue of testing vocabulary follows, which is ultimately the topic of this dissertation.

2.1.1 Corpus linguistics

By the late 1960s, the use of computers in every field of human activity was so widespread that those who created the first trial interface (BASEBALL) foretold that "many future computer-centred systems will require men to communicate with computers in natural language…" (Green *et al.* 1961:219) and, two decades later, Terry Winograd stated that "the computer shares with the human mind the ability to manipulate symbols and carry out complex processes that include making decisions on the basis of stored knowledge. […] Theoretical concepts of program and data can form the basis for building precise computational models of mental processing…" (1983:13).

As information technology progressed and appeared in several fields of human endeavour, it also 'invaded' the field of language testing and in 1996, Charles Alderson was the first to predict the potential use of corpora in language

assessment (cited in Taylor & Barker 2008:244). Less than ten years later, speakers at a symposium of the Language Testing Research Colloquium (Taylor *et al,* 2003 cited in Taylor & Barker, 2008: 244) discussed the use of corpora in the assessment of writing and reading.

Corpus (pl. Corpora) is a Latin word indicating a collection of linguistic data, selected and organized on the basis of explicit linguistic criteria in order to provide a sample of language (Beccaria 1996; De Mauro 2003; Sinclair 1996). Clearly, being a sample, a linguistic corpus cannot contain all the possible occurrences of the language, but an *a priori* choice of the kind of texts to be chosen must be made, so that the corpus is as close a statistically representative of the language as possible (Biber *et al,* 1998).

Corpus linguists distinguish different approaches to corpora: the corpus-based approach and the corpus-driven approach. In the former, analysis of the linguistic usage originates from a given theory or principle, or a particular linguistic trait in order to look for evidence which supports the theory within the corpus. In the corpus-driven approach, on the other hand, the starting point is to observe data, in order to formulate a theory based on such observations.

Today, it is possible to access via the Internet huge linguistic corpora such as the BNC (British National Corpus), or the CORIS (*Corpus di italiano scritto contemporaneo*[3]) (Rossini Favretti 2000), which contains 100 million words taken from oral and written language, from books, letters, dissertations and informal conversations of individuals of different age groups and with distinct social and geographical backgrounds (Bianco 2002). By the mid-nineties, corpora were being used in applied linguistics and in language pedagogy. Dictionaries like the Collins Cobuild English Language Dictionary were published (Vietri & Elia 1999) as well as grammars like the Longman Grammar of Spoken and Written English (Taylor & Barker 2008).

---

[3] Contemporary Written Italian Corpus – translator's note

Recently, research has developed in order to thoroughly investigate the lexical aspects of grammar with the use of specially designed software. Such software can carry out statistical and gloss functions. As known, a gloss (from the ancient Greek 'tongue' -- the organ -- as well as 'language') is a note made in the margins or between the lines of a book, in which the meaning of the text in its original language is explained, sometimes in another language. However, the "gloss function" in a database takes a parameter that represents a key in a glossary file and yields the resultant value, usually as a percentage.

Many projects have focused on lexical frequency which Alderson defines as "a crucial variable in text comprehension" (2007:383) and which is believed by many to be one of the main factors influencing performance in the comprehension, production, and the learning of language (Alekseev 1984; Geeraerts 1984; Muller 1979).

As will be explained later, the specific corpus developed for this study was used to create a list of words found in the two tests I analysed in terms of topic-specific vocabulary, mostly from military training and doctrine.

But first, I will describe the characteristics and typologies of corpora.

2.1.2 Corpora: characteristics and typologies.

Generally, corpora are of two types: closed corpora, which do not change, are usually text collections with a fixed size, and monitoring corpora to which it is possible to add or remove texts (open corpora). The latter is especially used in lexicographic studies on contemporary language In addition, a distinction can be made between native speaker corpora and learner corpora, consisting of texts produced by those who are acquiring a new language. Learner corpora provide useful empirical data for the systematic study of the learners' interlanguage (Alderson 1996). Granger (1988) claims that with Comparative Interlanguage Analysis (CIA) it is possible to identify both learners' errors and the "un-language" characteristics which can be identified through the over or under-use of particular words or expressions or idioms.

A corpus may simply be a collection of texts, or it might be enhanced by being annotated for the occurrence of various linguistic features through the use of special codes or tags which identify parts of speech. Such tagged or annotated corpora are a basis for further analysis (syntactical and semantic).

Although corpora were initially a tool of linguistic and lexicographic research (Biber *et al* 1998; McEnery & Wilson 1996; Sinclair 1991), the use of large amounts of text in electronic format (machine-readable form) has found application in several disciplines. For example, multilingual corpora containing texts belonging to two or more languages were developed.

There are also parallel corpora, comparable corpora, translation learner corpora and aligned corpora with a clear educational purpose as follows:

- Parallel corpora are useful to outline the strategies of professional translators (Pearson 2003);
- Comparable corpora provide information about the regularities within specific class or registers in different languages (Zanettin 1998);
- Translation learner corpora point out strategies and errors of learners, whilst fostering more awareness (Bowker & Bennison 2003); and
- Aligned corpora allow a valid confrontation of different translators.

Aligned corpora prove to be a particularly valid tool when studying contrastive lexical semantics, as in the case, for instance, of comparing how a particular situation is expressed in different languages and by different translators (Pearson 2003; Zanettin 1998).

2.1.3 Construction, tagging and coding of corpora

The construction of an electronic corpus is a rather complex procedure and quite difficult to summarize in a few lines. Briefly, a number of preliminary phases are involved, such as:

- the planning of the structure of the corpus,
- the acquisition of material (on paper, electronic or audio recorded),
- the breaking up of the boundaries of the lemmas (so-called tokenization),
- the rational distinction of lexemes and morphemes: in lexeme-based morphology, the derivation of meaning and the realization of phonological marking are distinct processes in word-formation. On the other hand, morpheme-based morphology assumes that language contains only one type of meaningful unit, the morpheme, which includes stems and affixes, all of which are signs (Aronoff 1994)
- The categories of words, verbs, nouns, adjectives, etc.
- the occurrences of textual words, the gloss function as mentioned earlier,
- the disambiguation of homographs (Chiari 2007:50-51).

Once the preliminary work is done, specific software is able to automatically produce concordances. This involves an alphabetical index of all the words in a text or corpus of texts, showing every contextual occurrence of a word, and identifying the more frequent clusters in a language. Data can be sorted according to Key Words in Context (KWIC) which consists of displaying all the occurrences of a word or syntagma (knot) in the centre of the computer screen, with a pre-determined number of words (collocates) to the right and left of the knot. The unit consisting of the knot and its collocates is called collocation.

Once the text collection is complete, in order for the corpus to become a source of linguistic data, it can be useful to annotate it with tags or markups; the tool used to assign such labels is called "Markup Language" (Gigliozzi 2003:73-77). Sinclair (1992:383) maintains that tagging is a fundamental operation because it shows the strict connection between form and meaning. Leech and his colleagues (1994:51) emphasized that there is no ideal way of tagging and Habert *et al.* (1997:48-53) discuss the various levels of annotation and the

associated difficulties. As Garside *et al* (1997:12) noted, a corpus can be tagged taking into account various linguistic and extra-linguistic factors, according to the degree of specificity of the information that must be provided, according to the nature of the data (written or oral) and to the styles of the text.

The above steps were followed in the creation of the corpus used in this research as will be described in the following chapters.

2.1.4 English for special purposes and the case of Military English

Everyday words are polysemous, in other words they can have more than one possible meaning; and that is the reason why they are useful, for they can be used in many situations. However, in some situations, such as in professional communication, everyday language may be too vague and not sufficiently specific.

Specialized language has developed among the members of a particular scientific or professional community, and from a lexical point of view, is characterized by the use of many technical terms.

When in 1968, the British Council organized the conference "Languages for Special Purposes", the acronym "LSP" spread very quickly. Ten years later though, the word 'Special' was changed to 'Specific' to mark the specificity of the linguistic needs of learners (Balboni 2000; Borello 1994; Gotti 1991).

According to Gotti (1991), in order for a language to be designated as specific, it should satisfy the following conditions: the emphasis on the user (didactic sphere), on the referent reality (pragmatic-functional sphere) and on the specialized use of the language (linguistic-professional sphere). These three conditions encompass the main aspects of a specialized language.

It was with this definition of language specificity in mind that the terminology of my corpus, was labelled as 'specific' or 'military-flavoured' on the one hand or, on the other, simply professional with no particular association with a specific

environment, as will be explained later. As Chung & Nation (2003:103) describe, specialized language can also be deemed as such thanks to 'an intuition of an expert', which is exactly one of the approaches I adopted to determine whether a specific term could be considered pertaining to the military environment or whether it was general enough to not belong to a category of texts in which prior knowledge of the topic would represent an advantage to some test takers and a hindrance to others.

In the following section, the testing of military English as implemented at the Italian Army Language School is described in order to better understand the rationale behind the research study.

2.2.1 Military English Testing

As mentioned in the introductory chapter, in the 1970s BILC developed levels of linguistic competence that derived from the rating scales of the US Interagency Language Roundtable (Green & Wall 2005, 379) which were subsequently adopted by NATO as "STANAG 6001: Language Proficiency Levels". Currently STANAG 6001 is the scale NATO countries use to define the linguistic requirements for personnel who are to be employed internationally; the scale is also used to adjust national procedures to international standards and as a basis for language testing procedures. As Green & Wall (2005:380-381) point out: "to qualify for posts within the Supreme Headquarters Allied Powers Europe (SHAPE), candidates would have to achieve the profile required by those posts […] Each PfP country has a specified number of posts within NATO for staff who, among other qualifications, often must meet certain STANAG levels of language proficiency".

It has been well established that the learning of a language and its assessment are interrelated, and the use of specialized language is of particular significance within this "relation" For a long time, the teaching of language for specific purposes (LSP) was focused almost exclusively on the acquisition of sector-based vocabulary. In the last twenty years, however, research has shown that

the actual specificity exists in the properties of the text (from which linguistic choices are made), in pragmatic factors (such as the addressee and his level of knowledge of the issue) as well as in the precision of word and concepts.

2.2.2 Background of research context

Since the end of the Cold War, foreign language training, and in particular the learning of English which is by far the most widely spread operative language, has become increasingly important for the Armed Forces of many nations. Each NATO nation has undertaken a huge commitment to standardize language tests with the assistance of agencies such as the Defence Language Institute in Monterey, California, the British Council Peacekeeping English Project and NATO's Bureau for International Language Cooperation (BILC). This demonstrates that politics plays an important role in many aspects of life and, as Alderson (2003) states, language assessment is no exception.

To better explain this last statement let us consider a multinational environment such as NATO or the UN: among the requisites a candidate must have to fulfil a position the knowledge of a language is paramount, be it English or another target language with the rationale that the higher the position, the higher the mastery of the language should be. Therefore, language assessment plays a fundamental role in deciding how certain key positions are assigned, given that a key position may very well not be assigned to a candidate and therefore to the country he/she represents due to his/her scores on the language test. Politics then.

In order to assess language knowledge in a standardized fashion, the Italian Defence has developed a multilevel proficiency test called Joint Forces Language Test (JFLT), which assesses linguistic competence in the four skills in adherence with the new descriptors of BILC's "Interpretation Document".

The reading and listening comprehension components of the test are common to the four Armed Forces (Army, Navy, Air Force and Military Police) and

contain geo-political topics and tasks common to all armed forces, whereas the writing and speaking components exploit different authentic situations which are more speciality-specific to the individual Armed Forces. For example, a member of the Italian Army may be asked to write a report on a specific peacekeeping situation whereas a member of the Italian Air Force may be asked to write or speak about a specific situation pertaining to his/her field of professional interest.

2.2.3 Stanag editions 1 and 2

Nevertheless, the descriptors in STANAG 6001 do not prescribe whether tests developed by NATO countries should be in general English or be ESP tests. In the case of military ESP, a particular duty at NATO requires a certain Standardized Language Profile (SLP), but it may be the case that the test development team members know very little about the tasks candidates are required to do with the language in a specific context.

As Green & Wall (2005:395) reported in their study "some teams have taken a general English approach in their testing, others have incorporated a 'military flavour', and still others have used texts taken from military sources and tasks based on military scenarios"

Those who prepare and validate English tests for military use are faced with numerous problems (Green & Wall, 2005:384), even though these are problems experienced by all sorts of ESP (Douglas 2001; Hamp-Lyons & Lumley 2001).

Indeed some of the emerging issues that have risen concern the linguistic competence and individual background of the testers themselves, affecting whether they are capable of developing an appropriate test for military settings (Bachman 1990; Davidson 1998; Lynch & Davidson 1997).

Other issues in military testing concern the coordination and harmonization of the several testing agencies that aim at enhancing standardization not only in

applying the STANAG scale, but also in evaluating the test results (Alderson *et al* 2004; Shohamy 2001).

These issues have been highlighted in a study carried out by Papageorgiu (2007:5-6) which involved the teaching of ESP to a group of military learners of English. The author reported that many test takers were expected to perform in the target language without any prior needs analysis of the situation carried out to determine which would have been the language tasks the test takers would have most likely encountered.

In her conclusion, the author claimed that the lack of a long tradition in the teaching of military English can easily result in what she calls the "Wild West of ESP" (2007: 15).

2.2.4 Test population

A few words must be spent on describing the peculiar Italian Army foreign language assessment agency located in Perugia, Italy. The school was established in 1965 to provide not only foreign language courses (lasting from one to four months depending on the typology) in main European languages, but also in many rare languages including, Pharsi, Urdu, etc. Besides offering courses, the school is the Italian national testing centre for all four Armed Forces.

Most Italian military personnel must renew their SLP every three years or in any case before deployment abroad. Personnel may sit for the proficiency test - the JFLT – which confers an SLP either as a student upon completion of a course, or as an external candidate following an official request. Needless to say, the JFLT is a very high stakes test for both candidates and stakeholders.

2.2.5 Prior knowledge of vocabulary and topics

The availability of a corpus based on military terminology might be a useful tool to identify vocabulary which is specific to work-related issues. Following Alderson's recommendations on using corpora in "language assessment including test writing, test construction, test scoring and score reporting" (1996 cited in Taylor & Barker 2008:245), the representativeness and relevance of the corpus resulting from the two tests I analysed were as carefully interpreted as their statistical analyses were. Just as Taylor & Barker (2008:249) predicted that even "small scale, specialized corpora…should not be underestimated as these can provide useful insights into task or domain-specific issues…", also my small scale corpus was influenced by the tasks test-takers were required to perform, either on work-related issues as on the TUI test or on broader professional issues as in the JFLT, the two tests described in the introductory chapter.

Vocabulary has long been recognized as one of the key components of L2 competence (Spolsky 1995); indeed the Test of English as a Foreign Language (TOEFL), which was established in 1964, dedicated an entire section to vocabulary (Read 2000; Schmitt 1999), and ten years later many studies pointed out the close relationship between acquired vocabulary, reading skill and comprehension of texts in L2 (Pike 1979). Spolsky however (1995:34) stated that vocabulary tests, in contrast to other types of evaluation, were more concerned with objectivity and reliability than with the validity of the way vocabulary was assessed.

Recent studies such as the one carried out by Read *(*2000:22*)*, have indicated the need for vocabulary tests to "require learners to perform tasks under contextual constraints that are relevant to the inferences to be made about their lexical ability". In his study Read underlines that one of the fundamental problems to solve in vocabulary tests is to bring vocabulary assessment in line with recent thinking in applied linguistics.

On the other hand, Stahl *et al* (1989) state that it could be argued that vocabulary knowledge on tests and prior knowledge are clearly linked. The

authors exemplify this by claiming that an expert in baseball is more likely to understand terms related to that sport which may very well be unfamiliar to non-experts of baseball. The authors further claim that test takers with prior knowledge of a topic will process content-specific terminology more quickly thanks to highly developed schemata. Moreover, inference i.e. a cognitive process used to build meaning (Hammadou 1991: 27) has long been shown to aid readers in understanding "even the simplest of texts" even if Afflerbach (1990 cited in Hammadou 1991:27) claims that only those readers who have prior knowledge of the topic use inference to understand the text.

Most studies on the role of inference have been conducted on L1 readers; little is known about if inference, prior knowledge and test scores are inter-related and if so, how and to what degree.

In fact, it is clear from the literature available on the issue of test takers' prior knowledge of the topic that military test takers' performance could be positively or negatively affected depending on the topic - and hence on the vocabulary it entails - presented in relation to their military training and current position within the Armed Forces.

The effect of vocabulary knowledge is important to take into account when interpreting the results of a reading comprehension test in which the texts may have been particularly relevant for a specific test population. According to Afflerbach (1990: 135 cited in Anderson & Freebody 1981; Spilich, *et al* 1979) foreign language readers may also access domain-specific vocabulary when accessing schemata.

Read (2000 : 190) reminds us that research findings have "well established that vocabulary is the most contributing factor - among many others of course- in reading comprehension of native speakers" and Laufer & Sim (1985 cited in Read 2000:190) confirm that even for non native readers, vocabulary was "what students needed most to make sense of what they were reading".

Although there is not a general agreement on the fact that some topics may entail knowledge of very precise vocabulary whereas others may include very broad vocabulary (Bugel & Buunk 1996:18), it is anyhow true that "a person who knows a great deal about a topic generally knows words specific to that topic" (Anderson & Freebody 1981 cited in Stahl *et al* 1989:30). This may very well cause bias in testing and represent a construct-irrelevant variance which threatens validity (Jennings *et al* 1999:428). These factors - which go beyond language ability - may be an advantage to some test takers and a disadvantage for others (Peretz & Shoham 1990: 447). This stance is in agreement with Alderson & Urquhart (1983 cited in Bachman 1990: 273) who found that students' test performance was just as affected by their knowledge of the content area as by their language proficiency. Bachman (1990:113) also suggests that factors such as educational and/or socio-economic background may very well affect test performance.

Peretz & Shoham (1990:448) claim that prior knowledge of a topic affects reading comprehension skills and that this effect is stronger and more noticeable in adult test takers since the latter tend to specialize in certain topics. Most people (Baldwin *et al* 1985 cited in Carrell 1998:286) have more knowledge about topics in which they are interested. This seems very relevant to the context of the participants in this study who are not only all adults from a common professional background, but also have proceeded to specialize in different military training, despite sharing common military doctrine.

The effect that topic has on test scores was investigated in an extensive study by Clapham (1996 cited in Douglas 2000) in which the performance on reading comprehension was correlated with the interest or background knowledge students had in the field or content area of the test. The findings indicated that if the content area was specific, test takers did better in their own subject area.

Clapham found these results to hold true for a found threshold below which students did not benefit from background knowledge. Although the effect or benefit did not increase with the proficiency level of the test taker, there was

anyhow a level above which benefit was gained from subject matter knowledge. These findings are consistent with Hammadou (1991 cited in Bugel & Buunk 1996: 17) in which "analysis of readers' inference did indirectly demonstrate that readers' background knowledge was affecting the comprehension process ….and that this is visible with the more proficient readers".

In their study, Jenning *et al* (1999) investigated the effect a topic-based test in an academic setting had on scores with an interest in measuring the advantages and disadvantages a test taker had in relation to his/her interest and prior knowledge of the topic. Although many studies (Anderson & Pearson 1984 cited in Afflerbach 1990) have been conducted on verifying how prior knowledge facilitates reading comprehension and most have concluded that there is a definite correlation between test scores and the effect of prior knowledge of the topics, Jenning *et al* (1999:430) argue that many of these effects "are highly dependent on the individual research methodologies". The results of some studies are contradictory according to Jenning *et al*, maybe due to the lack of a standard definition of "prior knowledge". Peretz & Shoham (1990:448) also claim that there is no agreed upon way to assess such knowledge.

Although the issue has been investigated thoroughly and it has been generally accepted (Bernhardt 1984; Johnson 1982, Peretz & Shoham 1990 cited in Bugel & Buunk 1996:17; Tan 1990) that background knowledge facilitates not only native readers' comprehension but also foreign language readers' skills; the effect of prior knowledge and vocabulary is often neglected when discussing reading comprehension skills. In fact, Papajohn (1999:72) who also found in his study on chemistry test scores that prior knowledge plays an important role, recommends further research in the precise role of topic in testing (1999:78).

2.2.6 Research Gap

Although extensive research has been conducted to investigate the effect that prior knowledge has on test performance, there is little available on this issue concerning the military environment. It is of great significance in a high stakes

test whose scores determine candidates' qualification for posts abroad or for career advancement. In this case it is particularly useful to investigate the impact that the introduction of a new high stakes proficiency test based on the descriptors of a revised proficiency scale (STANAG 6001 ed. 2) has had in terms of selected reading texts. The new reading comprehension component of the test includes texts from broader and less-military sources to reflect topical domains prescribed in the STANAG scales such as geo-politics, economy, culture, science and technology as well as from the professional field. Professional topics are not specified as such but may include reports, official correspondence, essays in which the reader must use language tasks such as understanding implicit information and writer's intent, learn through reading, understand abstract language used to support an opinion or used for argumentation whilst fulfilling the accuracy demands as per level three of STANAG 6001 ed. 2.

The role of prior knowledge in content areas familiar to the military test population has not been investigated for reading comprehension. The interaction between familiarity of topic due to prior professional or life experience and comprehension should be studied to see if this interaction indeed exists and to what extent it affects test takers' scores.

The validation of the JFLT could be at stakes if the inferences we can make on the scores of the reading comprehension component of the test are inaccurate for the scores could in reality reflect increased knowledge of specific topics. (This gap aroused my interest to further delve into the issue.) The contradicted qualitative studies carried out in which test takers' felt that military terminology actually helped them to understand the reading passage better. The issue of whether prior knowledge of the topic and of vocabulary consequentially specific to that topic, is fundamental to investigate as a possible explanation for this difference in score.

Therefore:

Does less specific military terminology of the new Joint Forces Language Test of the Italian Defence affect military personnel scores?

Chapter 3

METHODOLOGY

To investigate whether - and if so to what extent - a lower percentage of specific military terminology affected scores on the new test, I followed three separate chronological steps, each with different methodological approaches.

3.1   Building the corpus

As a reminder, the two proficiency tests – the TUI and the JFLT – are based on different editions of STANAG 6001: the former on edition 1 and consequently, the latter on edition 2. The second edition, entitled "The Interpretation Document" was annexed in 2001 to the original 1976 to provide additional insight in the shortcomings of the first edition. Specifically, the descriptors of the second edition are more detailed as to the content, task and accuracy demands required for language proficiency levels of 1, 2, 3 or 4[4] to be awarded. The most significant difference lies in the demands between a level two and a level three in that the requirements of the latter include being able to negotiate, analyse and argue about more abstract, geopolitical topics with occasional, non-patterned errors. Although most posts abroad require a certified level two, a level three is necessary for high profile posts such as military attachés or diplomatic positions entailing active participation in decision-making meetings or briefings.

Before selecting the sections of the two proficiency tests which would be administered to the sample test population, the full versions of both underwent scrutiny with a lexical analysis software. This procedure was necessary to be able to extrapolate different wordlists for the different analyses I will describe below.

First of all, I rapidly and easily examined a huge quantity of data from the two tests, which could then be organized in a clear fashion. Also, I worked with "Wordsmith tools", a software widely used for lexical analysis. A more detailed

---

[4]   Technically, the scale also comprises a level 5 (fully bilingual) which is, to my knowledge, not assessed within NATO countries.

description of this tool and how it was used in this research is given below. This specific study aimed at comparing technical military English and a more specific geopolitical language to which also military English belongs, especially at higher Stanag levels.

The starting point for my research was to create a list of key words I obtained from the lexical analysis software which analyzed the 5,000 words from the reading comprehension sections of both the first test (TUI, 1997) and the 10,000 words from the second test analyzed (JFLT, 2003). The reason the two tests differed in word size is related to the different topic concentration of the two tests: the TUI is the shorter of the two in that it concentrates on military doctrine and 'didactic issues'; there is little elaboration and extended discourse to understand supported opinion or point of view, hypothesis, implicit information, which are language functions prescribed in the second edition of the STANAG, on which the longer test is based. The texts on the TUI mostly deal with military topics with emphasis on doctrine and survival, emergency situations that typically occur abroad within the military context. On the other hand, the texts - at level three and four especially – of the reading comprehension section of the JFLT are on average longer since here they emphasize on extended and elaborate discourse of geopolitical, professional topics. This initial step was necessary to be able to create a list of potential military-specific terminology to guide the selection of the texts which would then be used in my research.

The second step I undertook was to use the corpus I had prepared in the months preceding the onset of my research.

Although I selected very carefully which texts to include in this corpus, i.e. military manuals, news bulletins, official statements from International Organizations, books and essays from the political and economic world, I did not include any markup language (Gigliozzi 2003:73-77), nor did I tag the tokens, since it was of no importance for the aim of my research. The texts were chosen from the sections of the two tests, which tested the topics and language

tasks listed in both editions of STANAG level three relating to professional and non-professional matters. I felt that level three better exemplified the use of military vs. non military terminology since the lower two levels of the test i.e. 1 and 2, do not concentrate on work/related issues *per se* but only on social, concrete and survival situations. Stanag level four would probably have been even more indicative for professional topics but it would have been virtually impossible to find students at that level at the time of this study.

Both tests are mini-versions of the actual proficiency tests administered at the Italian Defence. The full, original versions included 60 multiple-choice items whereas the mini tests now include only level three items amounting to a total of ten items per mini-test. The original TUI contained 13 level three items whereas the JFLT contained 15. I believed that twenty items (ten items times the two tests) would be a feasible and practical number to ask candidates to take time out of their courses or free time to sit for. The relatively small number of items would also allow for a better recollection of strategies adopted whilst answering the tasks during the subsequent interview.

The number of words of the newly devised mini-tests was now:
- mini- test TUI: 1,800 words
- mini test JFLT: 1,700 words

The opposite trend in number of words as compared to the full version is due to the fact that only 3 items were deleted from the level three section of the TUI test whereas 5 items were deleted from the level three section of the JFLT, as explained in the previous paragraph.

As a reminder, the items taken from the level three section of the reading comprehension component of the JFLT are considered classified material and could not be reproduced in an appendix as the TUI items are in appendix 7. However, for the aim of this dissertation it would be perhaps useful to compare the topics and the tasks of the two tests.

Table 1 Comparative table between topics and tasks of TUI and JFLT

Item #	Topic		Task	
	TUI	**JFLT**	**TUI**	**JFLT**
1	Paratroopers inside enemy lines	Agriculture (livestock)	Identify specific detail	Understand gist
2	Military doctrine/definition	Narration	Understand gist	Understand gist
3	Military doctrine	Narration	Understand gist	Understand gist
4	Military doctrine (part 1)	Intelligence ops	Understand gist	Understand gist
5	Military doctrine (part 2)	Politics	Identify supporting detail	Understand writer's intent
6	Military doctrine (part 3)	Politics	Identify minor detail	Understand Writer's attitude
7	Military doctrine (part 1)	Politics	Understand gist	Understand Implicit info
8	Military doctrine (part 2)	General unfamiliar issues	Identify supporting details	Inference
9	Correspondence on military issues (part 1)	General unfamiliar issues	Understand gist	Understand gist
10	Correspondence on military issues (part 2)	Immigration	Identify supporting detail	Understand implicit info

Table 2 Comparative table between text purpose and text type of TUI and JFLT

Item #	Text purpose		Text type	
	TUI	JFLT	TUI	JFLT
1	informative	informative	Professional matters	Economy
2	didactic	informative	Narrative	Economy
3	didactic	informative	Professional matters	General unfamiliar issues
4	didactic	evaluative	Professional matters	Professional issues
5	didactic	evaluative	Professional matters	Politics
6	didactic	evaluative	Professional matters	Editorial
7	didactic	evaluative	Professional matters	Editorial
8	didactic	evaluative	Professional matters	Editorial
9	informative	informative	Professional matters	Pamphlet
10	informative	evaluative	Professional matters	Editorial

As can be clearly seen from the table above, the texts of the mini version of the TUI mainly emphasizes military issues and doctrine with tasks that concentrate on understanding gist or identifying details. The purpose of the mini TUI text is to instruct or inform. The writer is anonymous. In the JFLT however, the text types are all professional matters with a voiced author who writes to provide an "evaluation" or rather an opinion or abstract elaboration of a topic.

Validity studies had been carried out on both tests and concurrent validity established between the two during the final validation procedures prior to the first official administration of the JFLT. However, for the aim of this dissertation, classical item analysis, descriptive statistics, and reliability coefficients were carried out only on the items of mini versions of both tests which I will from now on refer to as mini TUI and mini JFLT; the results of these analyses will be described in the next chapter.

3.1.2 *WordSmith tools*

The third step I undertook involved using the "WordSmith Tools" which is a lexical analysis software to analyze how words behave in texts. Among the different features the program offers, I used the "Wordlist", the "KeyWords" and the "Concord".

The "WordList" tool generates word lists that can be shown both in alphabetical and frequency order. These lists can be used to study the type of vocabulary used to identify common word clusters and to compare the frequency of a word in different text files or across genres.

In my research I will generate three different word lists, which derive from:

- a reference corpus of more than 4 million running words, which includes issues from the military field, geopolitics, law, geography and issues taken from international newspapers;
- the JFLT (Joint Forces Language Test), 10.579 running words;
- the TUI ("Test Unificato Interforze" or English Language Test), 5768 running words.

These texts, suitably elaborated, will then be used to generate Keywords.

The "KeyWords" function locates, identifies and analyses the words in the given texts. To do this, it compares the words in the mini tests (which will be respectively the JFLT and the TUI) with the reference set of words taken from

the larger corpus. Any word which is found to be outstanding in its frequency in the text is considered as "key". All words which appear in the shorter list are considered, and the key words are sorted according to its degree of outstandingness. If, for instance, the article "the" occurs 5% of the time in the JFLT wordlist and 6% of the time in the reference corpus, it will not be identified as a "key", although it may very well be the most frequent word. If the text concerns the description of a long-range missile, it may well turn out that the names of the constructor, and the words "explosive", "fuel", etc. may be more frequent than they would otherwise be in the reference corpus. To compute the "keyness" of an item, the program therefore calculates:

- its frequency in the short wordlist;
- the number of running words in the short wordlist;
- its frequency in the reference corpus;
- the number of running words in the reference corpus and cross-tabulates these.

This procedure was necessary to be able to classify the terms as 'military' or 'non-military'; this step would then make it possible to analyse the incidence of recurring specific or non-specific terms against the test scores to see if, how and to what extent, they differed in relation to the topic.

Subsequently, I began researching the terms within the three corpora. Each word was individually investigated according to its behaviour and use. This procedure was not new as something similar had been used during the development of the JFLT itself.

In fact, during the development of the Joint Forces Language Test, reference was constantly made to the British National Corpus to investigate whether the lexicon included in some of the more professional-specific texts were typically high or low frequency usages in actual social and professional areas. This not only offered insight into expectations of test–takers' knowledge of topic and vocabulary but also helped to choose authentic material.

The different word lists along with the quantitative information about the two tests would combine to help to answer my research question which is to investigate how, if and to what extent military terminology affects test takers' performance on a test. Therefore, once the percentage of military terminology was found on a test and ascertained how and if it belonged to a certain branch or specification, test takers' results were analysed in relation to the item and compared to the topical domain of the item.

3.2   Data Collection: Simple-group design

In order to test the hypothesis about the differences and relationships between the different score distribution of the two mini-tests – the TUI and the JFLT in relation to the terminology present in the two texts – I collected the data with a method that matched the research question I was investigating. So, in order to see whether the two sets of scores were correlated, I had the same group of individuals sit for the two tests. Subsequently, given that the sample was so small, I investigated the relationships between scores to make inferences using the small-sample t-test to validate the hypotheses about differences between the two means. Figure 1 shows which steps I conducted in my research design.

Figure 1 – Single-group design
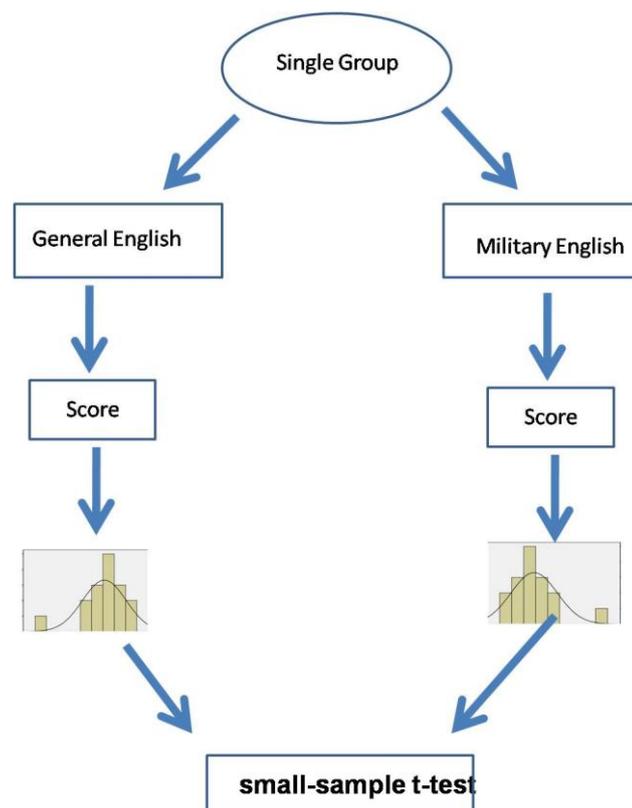
3.3 Test administration

I selected a group of 16 students who had been attending a three-month, refresher English course at the Italian Army Foreign Language School. I chose these particular students not only on the basis of their results on the diagnostic tests they had taken at regular intervals during the course they were attending, but also on the basis of tutors' assessment of their reading proficiency. However, I did not take into consideration which military branch or specialization they belonged to, nor if they had any prior experience serving in tours of duty abroad in international contexts. I chose those candidates who had already been assessed as a level 2+/3 in reading comprehension by their tutors. I approached these candidates and after having verbally informed them of my research project and design, asked them if they were willing to sit for a specifically-tailored reading comprehension test. Informed consent forms were signed and returned and the chain of command was informed. In addition, one of the candidates also accepted to give his contribution in a personal interview.

The two reading comprehension tests I specifically set up for the volunteers included ten items of the classified Italian Defence TUI level three reading comprehension component of the test and ten items from one of the two parallel (ascertained through extensive trialling) versions of the JFLT. At a later date, when these volunteers sat for the official administration of the JFLT which would confer them a SLP for qualification abroad, great care was taken to give them a different, parallel version. This was to avoid creating any advantageous conditions for them.

Only level three items were selected from both tests because it is only at this level that both editions of the STANAG prescribe specific military or professional-related topics. As previously stated, the main difference between the two proficiency tests is the topical range which is obvious - at a rapid glance - as being of more military flavour as in the case of the TUI and more general and/or geopolitical as in the case of the JFLT.

The administration of the test took place in standard conditions and was in compliance with the assessment procedures in use at the Army Language School. The students sat in a language laboratory with the purpose of the activity clearly explained to them. A consent form agreeing to take part in the research was signed and approved by the candidates' course director.

The reading comprehension tests were administered as follows:
– test-takers' identity cards were checked;
– the test administration procedure was explained;
– test takers' booklets and answer sheets were handed out;
– the reading comprehension test began.

For this research, no time slot was fixed. When test takers had completed the items, they handed in their booklets and answer sheets to the supervisor of the test session and only then were they allowed to leave the exam room.

Test takers read the items on the booklet that had been distributed. Each item was composed of a text, a stem and four options. They were informed both verbally and through written instructions[5], that there was only one correct answer per item. Also, they were informed that the results on this research would not affect their final scores on the proficiency test they were soon scheduled to take. After having read the text, test takers chose the correct answer and marked it on their answer sheet.

---

[5] The instructions in the TUI were in English; this original version was given to the volunteers. On the other hand, the JFLT has instructions in L1 and this version was given to the volunteers.

3.4   Interview

One of the 16 candidates volunteered to participate in an interview through which I hoped to gain additional insight into to see if, and to what extent, test takers were relying on known military terminology or prior knowledge (schemata) to answer the items on both tests. To do this, I asked the candidate if he would consent to being recorded on tape for later transcription and asked him to relate his thoughts in English. I made this decision because I feared that asking what is considered as being a minimally professional candidate (level three STANAG labelling) to verbally express his thoughts in L1 as he was taking a reading comprehension test in L2 would be doubly challenging for him. The candidate also agreed that this was the best procedure for him as well.

I asked him specific questions on what strategies he had adopted to answer each item. However, the candidate was somewhat reluctant to verbalize what he was doing so I found that I had to continually prompt and probe him to voice his thoughts. It is unclear whether having chosen to conduct this interview in English may have had an impact on the candidate's reluctancy. Finally, I asked for a global evaluation for both tests in terms of military topics being an advantage or on the contrary, a disadvantage. A detailed transcription of this interview can be found in Appendices 5 and 6.

The student (from now on referred to as Mr. "X") I invited to take part in this research agreed with interest. Mr. X had extensive professional experience, having worked several years in an international environment. His tutors indicated him as having the minimal professional competence to deal with authentic written material on professional topics; this is the minimal requirement to be evaluated as a threshold level three according to STANAG 6001, 2nd edition. Since Mr. X had these characteristics, I decided to submit both tests to him within a ten days' span, and began with the military-related test (TUI) firstly. Mister X was well aware he was being recorded and had duly signed a consent form beforehand. Both sessions lasted around 45 minutes.

Chapter 4

RESULTS AND DISCUSSION

In this chapter I shall analyze and discuss the data collected by means of the three methodological approaches described in the previous chapter. At the end of this chapter, I shall draw my conclusions on the basis of the data analyzed.

4.1 Corpora

In this section, I shall give an overview of the results of the analyses of wordlists, focussing on the most recurring words and providing examples of their occurrence within the test.

The created keywords lists located, identified and analysed the words in the given tests; these were created by comparing the words in the shorter tests with the reference set of words taken from the larger corpus. It can be assumed that key words give a reasonable indication of what the text is about. Therefore, any word deemed to be outstanding in its frequency in the text was considered "key".

As mentioned, the use of corpora made it possible to carry out a lexical study. In order to do this, I performed the following:

1. Initially I used three texts: the first was a collection of four million words I had personally devised over the course of many months. The sources I drew from included documents relating to: news articles from International press agencies, specialized military manuals, professional reports etc. My aim was to build a military terminology database. The second and third texts I used as sources were the two proficiency tests, the TUI and the JFLT.

2. Second, I used the software to create three separate wordlists. Each wordlist included the frequencies of occurrence of the tokens.

3. At this point, I created a list of "key words" by comparing the frequencies of occurrence in the shorter texts with the frequencies of occurrence of the

larger one. All the words in the shorter texts were analyzed. In order to find the 'keyness' the software computed:

- its frequency in the shorter text
- the total number of the words in the short wordlist
- its frequency in the larger corpus
- the total number of words in the larger corpus
- and cross-tabulated these using chi-square[6].

The following table summarizes my findings for the most relevant keyness in the mini TUI specific to my research:

---

[6] A test that uses the chi-square statistic to test the fit between a theoretical frequency distribution and a frequency distribution of observed data for which each observation may fall into one of several classes.

Table 3   TUI List of the first 30 key words

N	WORD	TUI (5.768 tokens)	REFERENCE TEXT (4.358.381 tokens)	KEYNESS
1	GOVERNMENT	3	5239	146,04
2	AGAINST	6	3973	90,02
3	POLITICAL	3	2964	74,09
4	STATE(S)	10	3784	68,0
5	PRESIDENT	3	2697	66,1
6	WAR	20	4403	53,1
7	GENERAL (rank)	8	2919	51,1
8	ARMY	14	3214	40,1
9	(WHITE) HOUSE	4	1855	36,1
10	LAW	4	1827	36,0
11	PARACHUTIST	4	3	32,1
12	LANGLEY	5	2	32,1
13	NATIONAL	7	1929	28,0
14	PEACE	6	12	26,1
15	SUPPORT	3	1282	24,1
16	LOGISTICS	5	152	26,0
17	THREAT	3	138	26,0
18	TROOPS	3	51	26,0
19	OPERATION	3	47	25,1
20	DOCTRINE	4	20	25,0
21	POLICY	4	18	25,0
22	STRATEGY	5	8	24,1
23	MILITARY	6	5	24,1
24	SECURITY	8	5	24,0
25	AIR FORCE	6	0	22,7
26	CHALLENGE	8	4	21,0
27	COMMITMENT	4	6	17,6
28	ORDER(S)	2	2	16,3
29	STABILITY	2	9	15,8
30	LEADER	3	6	15,4

I then analyzed these results to check for the exact use the key words had in the mini tests.

For example in the following output, the key word 'satellite' found in the mini JFLT occurs fourteen times; I analyzed each occurrence of the key word 'satellite' and the content of the item in which it occurred and the task or language function the test taker is called upon to perform. I believed the content domain was an important factor to verify as the topics usually reflect the vocabulary used. For the sake of brevity, I will not report each key word found to

pertain to military topics and invite the reader to refer to Table 5 further on for easy reference and to a more detailed analysis of item level data in section 4.3.

Table 4   Example of concordances (Wordsmith software output)

The frequency of key words in the mini TUI shows that 18 out of 30 terms can be considered military in nature (#s 6, 7, 8, 11, 14, 15, 16, 18, 19, 20, 22, 23, 24, 25, 28, 29, 30). The most recurring term in the TUI was 'war' with a frequency of 20 times (out of 5,000 words analyzed), followed by 'Army' (14 times), 'General' (as in military rank: 8 times).

These words presumably corresponded to military topics dealing with the items included in the mini TUI. The task students were asked to perform and the topical domain of the items are summarized in Table 1 on page 28.

As a reminder, the actual item and options cannot be displayed due to the fact that the mini JFLT is included in the larger official version of the test which is still protected as military classified material. However, Appendices 3 and 4 illustrate sample items included in the students' handbook made available to all test takers who are about to sit for the test. These examples can provide the reader with information on the tests and aid in understanding how the items and the tasks were developed according to the STANAG scale of proficiency levels.

The following table summarizes a list of key words that were created from the JFLT adopting the same procedure as the mini TUI explained previously.

Table 5   JFLT List of the first 30 key words.

N	WORD	JFLT (10.579 tokens)	REFERENCE TEXT (4.358.381 tokens)	KEYNESS
1	TECHNOLOGY	17	67	120,05
2	SATELLITE	14	47	103,02
3	UAV	9	4	92,04
4	SPACE	16	279	69,1
5	MILITARY	35	2308	69,0
6	SCIENTISTS	4	6	55,1
7	DNA	6	12	49,0
8	NEW	40	4292	48,0
9	OPIUM	9	97	47,0
10	DEPARTURES	5	5	46,0
11	AIRPORT	9	117	44,0
12	SALES	7	43	44,0
13	PROGRAMS	7	44	43,1
14	CANCELLED	6	22	43,0
15	SERVICE	14	495	42,1
16	TRADITIONAL	8	85	42,0
17	OVERSUPPLY	4	2	40,1
18	WARRANTY	5	11	40,0
19	RADAR	6	30	40,0
20	PEOPLE	34	3875	38,0
21	DEPRESSION	7	71	37,1
22	TICKETS	5	18	36,0
23	BASING	4	5	35,1
24	ARRIVALS	4	5	35,1
25	ONLINE	5	20	35,0
26	ELECTRONIC	6	47	35,0
27	BAND	8	138	35,0
28	BEACH	6	51	34,0
29	KURDISH	8	152	33,1
30	LASER	4	8	33,0

The table shows that for the mini JFLT only 5 words could be deemed 'military' (#s 3, 5, 15, 19, 29) either as such (that is to say, that the term could be used independently of a specific military topic as opposed to, for example, a term

such as 'UAV', unmanned aerial vehicle which definitely pertains to military situations) or as a prediction of a potentially military topic. Out of these, only 'military' stands out as being the second most recurring term (after a very neutral term: 'new') out of approximately 10,000 words.

In this research I decided to focus mainly on nouns and noun clauses and place less emphasis on other parts of speech such as verbs, pronoun, adjectives, etc. I felt nouns and noun clauses would be better indicators of the topic of the text and whether it had any relation to military issues in particular.

The following comparative table illustrates the ten items selected from the level three sections of the TUI and the JFLT – all pertaining to STANAG 6001 level three - in terms of topics and the terminology included in each item, which could be deemed 'military' or 'military-flavoured'.

Table 6   Terminology Comparison between JFLT and TUI

	JFLT		TUI	
N.	**TOPIC**	**TERMINOLOGY**	**TOPIC**	**TERMINOLOGY**
1	Economy	No military terminology	Narration (military situation: parachuting)	Enemy; parachutists; task; troops; to run for cover
2	Economy	Idem	Professional material (military definitions)	Friendly target; friendly fire; symbols
3	Narration	Idem	Professional material (military definition).	Armed reconnaissance; attacking targets
4	Geopolitics	Intelligence; Officer	Military doctrine	Army; strategies; threats; war
5	Editorial on the media	No military terminology	Idem	Peacetime engagement; deter conflict; hostilities; armed struggle
6	Geopolitics (on politicians)	Idem	Idem	Conventional forces; non combat; weaponry; application of force
7	Essay (excerpt on society)	Idem	Military doctrine	Trained; outfitted equipment; tailoring reserve forces; Cold War vestige; battlefield; coalition warfare
8	Narration	No military terminology	Idem	Readiness; force strength
9	Editorial on judicial systems	idem	Military correspondence	Corps engineer; trained and expertise; battalion -sized; units
10	News item on emigration	Asylum; refugees; threaten	Military correspondence	Idem

As can be clearly seen, there is an overwhelming majority of military topics and consequently terminology in the mini TUI - the construct of which was to test

understanding of the gist of military topics; the mastery of this specific lexicon could be assessed through items, which required candidates to understand reading passages clearly falling within military topical domains.

The items selected from the JFLT on the other hand, range from the very general to mainly geopolitical issues which contain a limited range of military vocabulary. Only item # 4 may deal with a military topic.

The task the reader had to perform in this item was mainly to understand the main idea and to identify writer's intent or implicit information contained in the text as illustrated in Table 4.

4.2   Descriptive Statistics

In this section, the descriptive statistics carried out on the two sets of ten items calculated with the software SPSS will be analyzed and discussed.

Table 7   Descriptive Statistics (*SPSS Output*)

Statistics

		JFLT	TUI
N	Valid	16	16
	Missing	0	0
Mean		6,63	4,88
Median		7,00	5,00
Mode		7	6
Std. Deviation		1,928	1,996
Skewness		-1,621	,081
Std. Error of Skewness		,564	,564
Kurtosis		4,153	-1,231
Std. Error of Kurtosis		1,091	1,091
Range		8	6
Minimum		1	2
Maximum		9	8
Percentiles	25	6,00	3,00
	50	7,00	5,00
	75	8,00	6,00

As illustrated in Table 7, when we analyze the relationship between the mode, the median and the mean we notice that in the mini JFLT test, these three values are very close with the median and mode having the same value (7) and the mean being slightly below (6.63). This indicates that the scores are closely clustered as these values are indicative of central tendency.

In the case of the mini TUI on the other hand, there is a completely different picture, with the three indicators of grouping certainly lower and slightly more spread out. It must be kept in mind however that since these scores are interval-scaled, the mean is the most appropriate indicator of central tendency (Bachman 2004: 63). This is a first possible indicator of a noticeable difference between the two tests; test-takers scored higher on the mini JFLT than on the mini TUI.

Score dispersion is analyzed through its three principal indicators: range, semi-interquartile range and standard deviation. The range is very wide for both tests but this still does not add much to the analysis, as there could be simply two candidates for each test who reached very high or very low scores (outliers). The interquartile range on the other hand indicates the variability that is based on the range of the middle 50 per cent of the test scores. According to Bachman (2004: 64), the semi-interquartile range is useful with highly skewed distributions, so this is not our case as the skewness of both tests falls within the rule of thumb range of +2, -2.

Figure 2.   Histogram mini JFLT (*SPSS Output*)
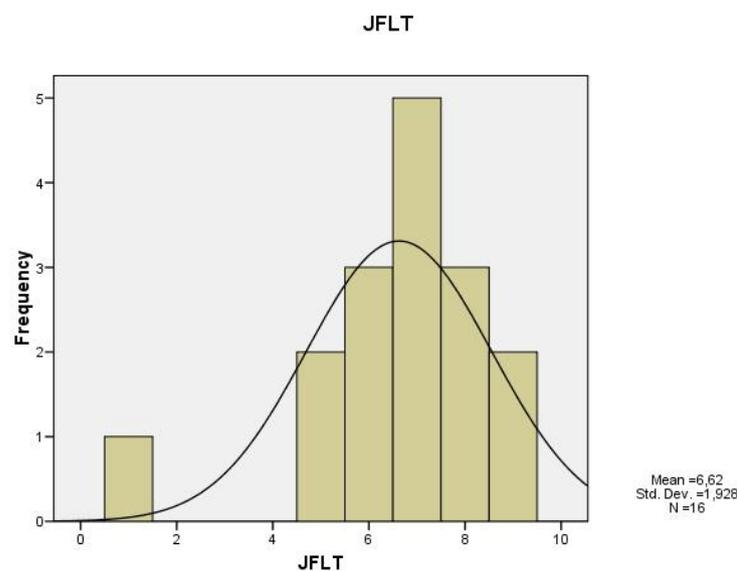
Figure 2 shows the score distribution of the mini JFLT test. The shape of the score distribution is slightly peaked, with a kurtosis value of 4.15 indicating a non-normal distribution.

The skewness is negative (-1.62) but within the range, with the distribution being a bit off-centred to the right. In such a negatively skewed distribution, high scores have the highest frequency. The value of the mean (6.62) is lower than

the median (7) and the mode (7) although this could be just due to extreme scores affecting the mean.

Figure 3. Histogram mini TUI (*SPSS Output*)
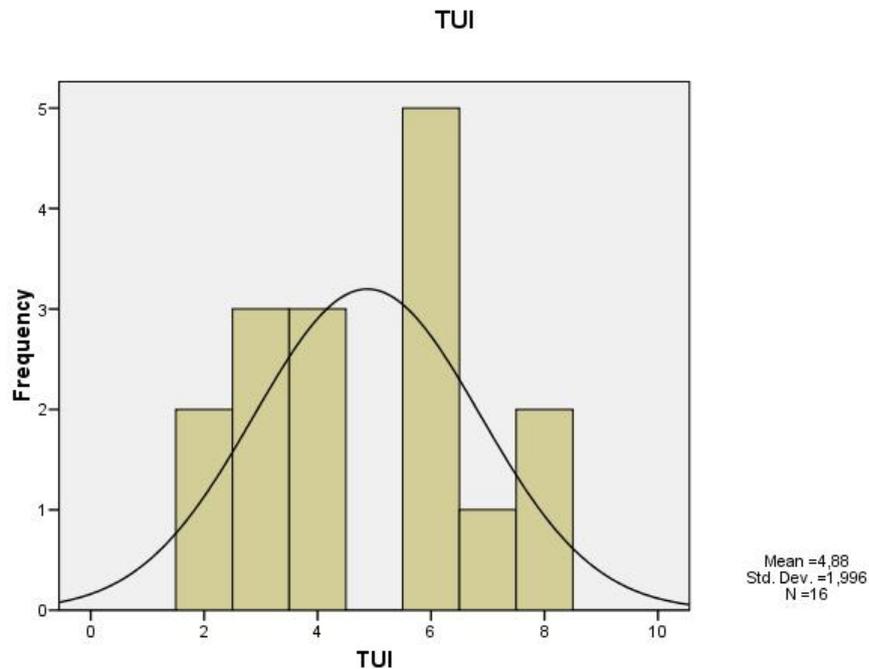
Figure 3 illustrates the score distribution of the mini TUI scores. It can be noticed that the distribution is bell-shaped with the kurtosis value within the normal range (-1.231). The value of skewness is close to zero (0.81) with the distribution being quite off-centre to the left. Also in this case the value of the mean (4.88) is smaller than the median (5) and the mode (6).

Therefore, based on the distribution of scores, candidates found the mini TUI test more difficult than the mini JFLT.

4.3   Classical Item Analysis

Item analysis was carried out to check for the distribution of the test scores to see how the tests were perceived in terms of difficulty and to see which items were failing to function (Bachman 2004: 121). The first part of this item analysis (IA) focuses on the item difficulty (hereon referred to as Facility Value), that is the proportion of test takers who answered the item correctly. One of the main limitations of IA is that it looks at only one aspect of the procedure, i.e. the item (Bachman 2004: 141). Also, it must be borne in mind that IA is strictly sample-based and may likely be different in another test population analysis (Bachman 2004:139).

Table 8: mini TUI facility values

mini TUI: FACILITY VALUES

ITEM NUMBER	1	2	3	4	5	6	7	8	9	10
FACILITY VALUE (FV)	100%	56%	63%	25%	31%	63%	44%	56%	25%	25%

At a first glance at the table above, we can see that in the TUI test, with the exception of item 1 that has a FV of 100%, indicating that students found this item extremely easy to answer, the majority of the remaining items have from average to low FVs (less than 50%). Given that the selected sample of students was known to be at level three reading proficiency as indicated by diagnostic scoring and tutors' in-class assessment, an explanation of this easy item #1 could indicate that the correct answer is clearly obvious and that the other distracters are unattractive (Bachman 2004:137). On the contrary, the other items show a wide range of difficulty as perceived by test takers. Indeed, whereas items #s 2, 3, 6, 7, 8 could be deemed acceptable in terms of difficulty, items #s 4, 5, 9, 10 are clearly too difficult.

Specifically, item # 4 is one of the two items which assess comprehension of supporting detail pertaining to the same text on military doctrine. As seen in Table 9 showing distracter analysis, the correct answer D in item 4 was chosen by fewer test takers than the more attractive option C. It could be argued that the difficulty of this item lays precisely in the ambiguity of option C which attracts lower achievers who are influenced by common military knowledge and NOT according to what the text states, and that is, that the strategic environment is linked to appropriate military actions and is not 'the consequence of a variety of responses,' precisely as option C states. Therefore, since the discrimination index is quite acceptable at .40, lower achievers could be applying their own knowledge of the topic perhaps as learned throughout courses at the military academy or through acquired experience, instead of applying what was read and asked for in the item.

Also for item # 5, option B attracts much more than the key (C). Once again, given the acceptable discrimination index, lower achievers may very well be applying their prior knowledge of the military topic of peacetime engagement to answer the item instead of answering according to what is stated in the text.

Finally, items 9 and 10 with very low facility values of 25% and discrimination indices respectively at .20 (low) and .80 (very good), both refer to the same text which is an example of official correspondence regarding the military topic of a corps engineer training request. Options A and C actually attract more than the key B even among high achievers (given its minimally acceptable discrimination index of .20). The task is to understand inference (the key states that the General 'means to explain the skills, abilities and training which are needed by a corps engineer'. Although this can be easily inferred from the details which support the sentence in the text 'I wish to define groups of tasks at which to aim training'), this item still creates a problem even among the better students. Arguably, option A could very well be in line with normal procedures Army generals are supposed to adhere to which is a fact that test takers are familiar with. Once again then, test takers seem to apply common knowledge rather than perform the task they are called upon to do.

Item 10 is fairly difficult as well since its facility value is fairly low at 25% however it discriminates very nicely at .80. Although options A and C attract just as much or even more than the key, the percentages of high achievers getting the correct answer is greater than those of low achievers.

Table 9: mini TUI Distribution of distracters

mini TUI DISTRIBUTION OF DISTRACTERS

	1	2	3	4	5	6	7	8	9	10
A	0%	25%	25%	31%	13%	6%	38%	0%	31%	25%
B	100%	6%	6%	6%	50%	31%	44%	38%	25%	25%
C	0%	56%	6%	38%	31%	63%	6%	6%	38%	38%
D	0%	13%	63%	25%	6%	0%	13%	56%	6%	13%

N.B.   key = green;

option not chosen = red;

Table 10: mini TUI discrimination index

Mini TUI: DISCRIMINATION INDICES

Item	1	2	3	4	5	6	7	8	9	10
DISCRIMINATION INDEX	0.00	1.00	0.40	0.40	0.60	0.40	0.20	0.40	0.20	0.80

Table 10 above shows that, apart from item 1 that would obviously not discriminate as all 16 candidates got the item correct, and item # 7 which has a DI slightly under the recommended .30 (Bachman 2004: 138), all the others have a greater value, with # 2 item having a perfect DI.

Table 11: mini JFLT facility values

mini JFLT: FACILITY VALUES

ITEM NUMBER	1	2	3	4	5	6	7	8	9	10
FACILITY VALUE (FV)	56%	69%	94%	88%	81%	69%	38%	63%	50%	56%

Table 11 above shows the facility values for the JFLT items; we can notice that items 3, 4 and 5 proved to be easy for candidates as almost all of them chose the right answer, whereas the other items are within acceptable FV limits with the exception of perhaps # 7 which is the most difficult item for this sample test population.

As a reminder, item # 4 is the sole item referring to a geopolitical/military topic involving an intelligence officer. When asked about this item, the sole volunteer I interviewed stated that he had to read the text twice and deduce the correct answer (Appendix 6). As mentioned, although the actual content of the items cannot be disclosed, the topic of the text in this item is geopolitical given that it talks about how language is used in politics. The task the candidates are to perform was to identify the writer's attitude (the stem reads…'the writer's attitude clearly reflects that…'). Option A attracted more than the key although its discrimination index is high at .80. It could very well be argued that low achievers were attracted to option A because the content of that option is mentioned in the text and although not entirely reflecting the writer's attitude, it could nevertheless be seen as a supporting detail of his/her attitude. Contrarily, high achievers are not distracted by this nuance.

Table 12 below illustrates that, apart from the already noticed item 1, items 6 and 8 include distracters which test takers deemed not plausible enough to be chosen. If analyzed in connection to their facility values, it is clear to see how especially in the case of items 4, 6, 9 and 10, the distracters were distracting candidates too much from the correct answer. This information will be better explained and described in the interview I conducted with one test taker during which I probed into the reasons some options were chosen over others and what was in the text. This was especially in relation to military topics and terminology, which may have had an impact on the test taker's choice of option.

Table 12: mini JFLT Distribution of distracters

mini JFLT DISTRIBUTION OF DISTRACTERS

	1	2	3	4	5	6	7	8	9	10
A	31%	6%	0%	88%	6%	0%	44%	25%	31%	13%
B	56%	19%	94%	6%	6%	13%	13%	63%	50%	6%
C	13%	6%	0%	0%	0%	13%	0%	0%	6%	56%
D	0%	69%	6%	0%	81%	69%	38%	6%	6%	19%
?	0%	0%	0%	6%	6%	6%	6%	6%	6%	6%

N.B.   key = green;

option not chosen = red;

In Table 12 above, we can notice that, apart from the already singled out item #s 3, 4 and 5 which presented high FVs, also items 1, 5, 6, 7 and 8 included one distracter that did not entice candidates enough to be chosen. Once again, the interview will shed light on the factors that led the test taker to ignore some of the distracters and choose the option he did.

Table 13: mini JFLT: discrimination index

Mini JFLT: DISCRIMINATION INDICES

item	1	2	3	4	5	6	7	8	9	10
DISCRIMINATION INDEX	0.00	0.20	0.20	0.40	0.40	0.60	0.80	0.40	0.40	0.40

Table 13 shows how these items do not discriminate well between high and low achievers. Specifically item 1 does not discriminate at all although its facility value is fairly acceptable at 56% Items # 2 and 3 have a value slightly below the desired .30 and the others discriminate at various degrees between .40 and .80. On the whole, it would seem that options in the mini TUI functioned better that those in the mini JFLT; in fact, there were more options in the mini JFLT which were left unchosen as compared to the mini TUI whose options were all chosen to some degree. This could imply two possible reasons:

- there are more ambiguous options in the TUI than in the mini JFLT although the discrimination indices would not seem to support this;
- test takers make more of a conscious effort to tackle each option as a possible key and make a reasoned choice based on a careful reading of the text.

4.4 Reliability of the mini tests

Studies of reliability that were carried out showed that coefficient Alpha was very low: this was to be expected as the test was too short.

Table 14   TUI and JFLT Reliability Statistics (*SPSS Output*)

JFLT - Reliability Statistics

Cronbach's Alpha	N of Items
,486	10

TUI - Reliability Statistics

Cronbach's Alpha	N of Items
,521	10

As test length can affect the reliability of a test, it came as no surprise that the reliability indices of the mini tests were rather low; therefore I decided to calculate Spearman-Brown's Prophecy formula which is used to estimate the reliability of a longer test. The following formula is a general form of the Spearman-Brown correction for length which assumes that "the additional items in the test would have the same reliability as the ones already in the test" (Bachman 2004:164).

Table 15   JFLT Spearman-Brown Prophecy

JFLT

SPEARMAN-BROWN PROPHECY $r_{xx'} = \dfrac{kr_{hh'}}{1+((k-1)r_{hh'})}$

NUMBER OF QUESTIONS	K: The factor by which you would lenghten the test	CRONBACH'S ALPHA ($r_{hh'}$) = .486	RELIABILITY COEFFICIENT AS A RESUL OF THE ABOVE SPEARMAN-BROWN PROPHECY FORMULA
10	1	0,486	0,49
20	2	0,486	0,65
30	3	0,486	0,74
40	4	0,486	0,79
41	4,1	0,486	0,79
42	4,2	0,486	0,80

As can be seen above, 42 items on the JFLT would be needed to reach a reliability index of .80, deemed acceptable.

Table 16   TUI Spearman-Brown Prophecy

TUI

SPEARMAN-BROWN PROPHECY $r_{xx'} = \dfrac{kr_{hh'}}{1+((k-1)r_{hh'})}$

NUMBER OF QUESTIONS	K: The factor by which you would lenghten the test	CRONBACH'S ALPHA ($r_{hh'}$) = .521	RELIABILITY COEFFICIENT AS A RESUL OF THE ABOVE SPEARMAN-BROWN PROPHECY FORMULA
10	1	0,521	0,52
20	2	0,521	0,69
30	3	0,521	0,77
31	3,1	0,521	0,77
32	3,2	0,521	0,78
33	3,3	0,521	0,78
34	3,4	0,521	0,79
35	3,5	0,521	0,79
36	3,6	0,521	0,80

On the other hand, only 36 items would be needed on the TUI to reach the same reliability index of .80.

4.5 Correlation between results

Correlation is the core of many approaches for investigating both reliability and validity of measurement. A correlation coefficient is a statistic that is calculated from data that summarizes the strength and direction of the relationship between two variables. In our case, it is not appropriate to use and interpret the Pearson product-moment correlation coefficient as one of the three assumptions about the scores that the candidates achieved on the mini JFLT and the mini TUI has been violated (Bachman 2004: 85). Specifically the scores on the mini JFLT are not normally distributed. (Therefore, it was more appropriate to calculate the Spearman correlation coefficient as follows:

Table 17:   Spearman Correlation between scores on the mini JFLT and the mini TUI (*SPSS Output*)

Correlations

			JFLT	TUI
Spearman's rho	JFLT	Correlation Coefficient	1,000	,433
		Sig. (2-tailed)	.	,094
		N	16	16
	TUI	Correlation Coefficient	,433	1,000
		Sig. (2-tailed)	,094	.
		N	16	16

The table above shows the SPSS output for the correlation value r = ,433. However, this was not significant (p=.094) and therefore there is no relationship between the two tests.

4.6 T-test – Single Group Design

The Paired-Samples T Test procedure was used to test the hypothesis of no difference between two variables. The data may consist of two measurements taken on the same subject or one measurement taken on a matched pair of subjects: in our case, the test scores of the mini TUI and mini JFLT.

Additionally, the procedure produces:

- descriptive statistics for each test variable;

- a confidence interval for the average difference (95% differently specified value).

Having rejected the null hypothesis, which is:

- the means for JFLT and TUI ratings will be equal: (H$_0$: $\overline{X}_{JFLT} = \overline{X}_{TUI}$),

we can now restate our research hypothesis:

- the means for JFLT and TUI ratings will NOT be equal (H$_1$: $\overline{X}_{JFLT} \neq \overline{X}_{TUI}$).

Considering the fact that the same testing population sat for both tests (single-group design), the t-test will be a dependent one.

I used a 95% Confidence Interval, because on the basis of the results no decisions other than testing a research hypothesis will be made. If the results had affected decisions about programs or the assessment, then a 99% Confidence Interval would have been more appropriate (Bachman 2004: 173)

In addition, given that the aim of the research is to investigate how the incidence of less specific military terminology has affected scores on the new test, the research hypothesis will be stated as non-directional since it was not yet known whether the mean of the JFLT was larger or lower than the mean of the TUI.

Table 18   Paired Samples Statistics (*SPSS Output*)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	JFLT	66,25	16	19,279	4,820
	TUI	48,75	16	19,958	4,990

The first output of SPSS, in Table 16, shows that for all 16 subjects, the mini TUI mean scores differed, on average by 17.5% in comparison to the mini JFLT scores.

The table below shows the final output.

Table 19   Paired Samples Test (*SPSS Output*)

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	JFLT - TUI	17,500	11,255	2,814	11,503	23,497	6,220	15	,000

The mean column in the paired-samples t test table displays the average difference between the mini JFLT and mini TUI measurements. The standard deviation displays the standard deviation of the average difference score. The Standard Error Mean column provides an index of the variability one can expect in repeated random samples of 16 students similar to the ones in this study. The 95% confidence interval of the difference provides an estimate of the boundaries between which the true mean difference lies in 95% of all possible random samples of 16 test takers similar to the ones participating in this study.

The Sig. (2-tailed) column displays a statistically significant probability of obtaining a t statistic whose absolute value is equal to or greater than the obtained t statistic.

Since the significance value for the different scores obtained on the two tests is less than 0.05, we can conclude that the two tests are significantly different from each other.

4.7   Interview feedback

During the interview, Mr. X had interesting insights and after my first attempts to probe into the strategies he had adopted to deal with the test items he continued to relate to me his considerations whilst answering the items, albeit with some difficulty and often reluctantly. Even though more than once I emphasised that he could take all the time he needed to answer my questions, it was clearly perceivable that he was struggling to try to finish as soon as possible.

The strategy that seemed to emerge from his feedback was that he usually read the entire text first, then read the options and finally scanned the text again in search of key words, which could reveal the right answer to him. Only in a few circumstances did he decide to read the options ahead and then read the text, but he was never content with this approach and shortly after he returned to his original strategy.

The most important elements which emerged from the interview (Appendix 5) are the candidate's comments on the individual items especially as pertaining to the mini TUI. In fact, he commented that he had gone straight for the right answer as the definition of armed reconnaissance was very clear. Although information was not collected on volunteers' professional background, basic definitions of terms such as an armed reconnaissance, is common knowledge to all military personnel. This example is very similar to what follows in the interview when he also states that on item 5 on the mini TUI, the correct answer was easy to find even if the text did not say directly, but '....it is clear that you've got to avoid local strife...".

The most remarkable thing that the candidate mentioned soon after taking the first test i.e. the mini TUI, was his appreciation for the test, saying that he preferred to deal with more familiar situations as opposed to more general topics. He did however, express doubts as to whether this would constitute an advantage or rather create a false sense of confidence. This is perhaps what

actually did happen in that his results on the tests proved that he scored very highly on the mini JFLT and poorly on the military one (mini TUI).

His comments on the mini JFLT seem to further confirm his perception of added difficulty on items which dealt with general knowledge. He confessed that he had to make his own evaluation (Appendix 6:89) whilst answering the tasks and that at times, he had to read the text twice. His overall evaluation of both tests was that the mini TUI could give him more chances to get the right answer because it deals with military topics, although admitting that the main issue is dealing with time constraints.

On the basis of the above and considering that volunteers were not given time limitations to finish the twenty items, it can be concluded that perhaps a false sense of security yielded by familiar topics reduces concentration and affects test performance.

4.8   Overall Results

Corpus analysis showed that the mini TUI had a very strong percentage of military or job-related terminology in comparison to the mini JFLT which had a wider range of topics including economics, culture, science and technology as well as military topics intended for the general reader. The "key-word" feature in the "Wordsmith tools" software, by comparing the words in the smaller sized tests (in turn the mini JFLT and the mini TUI) with the reference set of words taken from a larger four-million word corpus, found those words outstanding in their frequency in the text and tagged them as "key". These words were finally sorted in order of outstandingness. The outputs (tables 3 and 5) showed how the hypothesis, i.e. that the difference in the frequency of military topics and therefore of fewer specific terminology would affect test takers' performance, without any doubts.

Being aware that the sample of students all possessed the same level of competency, it can be affirmed that the mini TUI test was more difficult than the JFLT since the average mean score of the latter was 17.5% higher than the

former. This was revealed both by descriptive and distribution statistics but also through triangulation with the interview I conducted with a volunteer to ascertain the results of the two tests.

The interview conducted on both tests mostly showed how test-takers may have a false sense of security believing that they are sitting a test which will deal with (supposedly) familiar topics. It could be then argued that this false sense of confidence was in some way what contributed to affect the final results. Overall, it can be concluded from this small sample investigation that a high incidence of military terminology in a proficiency test affects test-takers' performance and that test-takers may very well approach the test with a false sense of confidence.

Contrarily, it could be also argued that, these topics may be biased towards some military personnel belonging to specific branches or specialty as opposed to others depending on their professional backgrounds.

However, in my limited research, there is a significant difference in performance on the mini TUI and the mini JFLT most probably in relation to the incidence of military terminology. This can be arguably be attributed to:

- the greater incidence of military topics in the mini TUI and its lower mean;
- the virtually non-existent military topics in the mini JFLT and its higher mean;
- the findings of the keyness incidence for both tests in relation to the specifically- created corpus;
- triangulation with the interview findings during which the sole volunteer does hint at military topics creating a false sense of confidence with the item luring the test taker into producing an answer based on his world knowledge of the topic and not on what the text is assessing.

The difference in test takers' performance on the test is in fact confirmed in the interview whereby the candidate revealed how he relied on specific prior knowledge of topics and lexicon to guide towards the key. Unmistakably this led him in the wrong direction as it did for the other 15 volunteers of this study.

Therefore, the answer to my research question:

"Does less specific military terminology of the new Joint Forces Language test of the Italian Defence affect military personnel scores?"

can reliably be answered on a basis of the results obtained as such:

In a test with a high frequency of military related terminology, the scores are negatively affected.

Chapter 5

LIMITATIONS, DISCUSSION AND FUTURE RESEARCH

In this chapter I will highlight the limitations of the study I carried out, followed by a discussion of the results found to answer my research question and future research I hope can be dedicated to this topic.

5.1 Limitations

Despite extensive data being gathered to carry out this research, there are nevertheless some limitations. The main limitation of this study is linked to the small sample available; indeed, sixteen test-takers is not an extremely statistically reliable number if important decisions are to be taken regarding the implementation of the military proficiency test, considered to be a very high-stakes exam. A number ranging between 150 and 200 students should be necessary for a more in-depth and reliable analysis. A larger number of students would reveal if my findings are consistently true or if they are linked to this small sample.

Another limitation concerns the qualitative data-gathering I adopted. By having benefited from the interview from just one informant, the ensuing result that the military-flavoured test is harder than the other may not be very reliable or consistent. This finding could very well be just a subjective judgment of one candidate on a selected range of items. The triangulation with the use of a questionnaire at the end of the test session to be handed out to test takers could have probably helped to shed further light on this issue.

Also, it could be argued that the very selection of the items representing STANAG editions 1 and 2 of level three reading comprehension should have been perhaps analyzed by a pool of experts who could judge the level, content, and tasks of each item against the STANAG scale and correlate them firstly within the individual test and then between the two. More items covering a larger representative sample of topics of the level and perhaps analysed in

relation and correlation to the military background of the specific test takers could have provided more information. Also items could have been selected from the other skills of listening, speaking and writing as well to check for a trend across skills.

Within these limitations, the following emerged from my small scaled study.

5.2 Research questions

The main purpose of this research was to investigate if and to what extent the prevalence of military topics could affect test takers' performance either positively, i.e. familiarity with the topic thanks to prior military training or work experience which provided help in answering the items regardless of the language function being assessed or, on the contrary, negatively, i.e. specific topic might create a false sense of confidence that induces test takers to respond without carefully reading the task they are asked to perform. This study was specifically designed on two proficiency tests developed at the Italian defence: the official English Proficiency test – JFLT and the test it replaced in 2006 – the TUI.

To find the answers to my research questions, a corpus study was carried out in which the keyness of military words, assumed to belong to military-specific topics was created and analyzed. An overwhelmingly amount of military terminology was found in the TUI whereas one recurring word only which could be deemed military in nature, was found for the JFLT.

A selection of ten level three reading comprehension items from both tests were arranged in booklet form and administered to 16 volunteers. Item analyses and descriptive statistics revealed which items resulted easy or difficult for this small sample of test takers. These results along with item level data showed that the TUI was more difficult for these test takers.

These results were triangulated with the outcome of an interview conducted on one candidate which provided only limited useful information; among some of the information was that the TUI items were easier to respond given that the topic was familiar whereas the JFLT items seemed trickier due to the fact that the topics were unfamiliar. In fact, the candidate suggested that perhaps this could create a sense of confidence that could lead the test takers to respond without reading.

A paired-sample t-test was carried out; results showed that the difference in means of the two tests on the same population was not due to chance and that there was indeed a relationship between the two scores.

It would appear that, given the keyness found in the two corpora developed, the higher percentage of military topics in the TUI and its lower mean, the meaningful relationship found between the means and the triangulation with the candidate's comments made during the interview, the familiarity with military topics on a reading comprehension test does indeed affect test takers' performance negatively in that it gives a false sense of confidence.

5.3 Future research

Given the fact that the JFLT is a very high stakes test the scores of which not only affect students' careers but also guide stakeholders' decisions on whether the inferences we make on the language ability to perform in international scenarios ranging from the diplomatic to military-civilian relations to life-threatening situations are correct, I believe further research in the suitability of a valid and reliable assessment tool is paramount.

The validity of the JFLT was established before its official administration in 2006. Besides the procedures followed were in conformity with the typical test development phase and that is moderation, revision and initial pre testing on ten British officers at SHAPE. Further steps to establish the validity of the JFLT were then followed by trialling items per level on candidates who were known to be at that level either thanks to tutors' in-class assessment or through

diagnostic assessments. Statistical studies provided information on which items were to be discarded and which were to be revised and kept; finally the items were arranged in booklet form and trialled on a group of 90 Staff Officers who were asked to volunteer to trial the test close to the end of a three-month, 34-hour a week English course the objective of which is to achieve straight level 3 on the STANAG 6001 ed. 2 scale. The timing was not set by chance as they would be sitting for their final proficiency test after a week so that correlation studies could be carried out between the scores of the JFLT and the TUI. The staff officer course is highly demanding; a Standardized Language Profile of 3-3-3-3 is a prerequisite for these officers to qualify for a one year training course at the General Staff. This career path will lead them to become one of the few decision-making generals of the Army.

Once again, statistical analyses and qualitative analyses methods were carried out to establish the validity and reliability of the JFLT. Correlation studies with the TUI were performed to check for relationship between the two scores. However, concurrent validity established with the declassified TUI may not be sufficient. As a reminder, the TUI was a proficiency test based on the description of STANAG 6001 edition 1. The main differences between edition 1 and 2 are a qualitative leap between levels 2 and 3 in terms of content, task and accuracy. Therefore, further concurrent validity with perhaps an external testing tool based on STANAG edition 2 is necessary.

Indeed, the Benchmark Advisory Test (BAT), developed by a selected working group within BILC, is a multilevel proficiency test which assesses the skills of listening, speaking, reading, and writing in compliance with the descriptors of STANAG 6001 ed. 2 items dealing with representative topical ranges as per the Stanag levels with tasks reflecting the requirements of each level. This test will not replace national tests but will serve only to establish concurrent validity with existing national tests the results of which will be compared against the BAT scores. The reading, listening and writing components of the test will be Computer-delivered and scored, whereas the speaking test will be a telephonic Oral Proficiency Interview carried out by trained raters and testers. Levels will be reported in the range from 0 to 3. Upon validation, the items will be posted to

a secure website and electronically delivered in the presence of a Protocol Officer in the period between February and September 2009. BAT administrations will take place within the nations involved in the project who have expressed an interest in this test. At the moment, only twenty administrations of this test have been allocated to Italy and these are being conducted from June to September 2009.

Comparative studies between the BAT and NATO national tests would also enhance the ultimate goal of the STANAG scale which is to provide standardized language proficiency levels as equally and mutually intended by all participating member nations.

These studies directed also towards the investigation of corpora on the two tests would further enhance or on the contrary, dispute the findings of my research. In any case, more noteworthy and reliable data will be provided as to what impact specific military terminology has on military test-takers' performance not only on a reading comprehension test but on a proficiency test as a whole.

REFERENCES

1. Afflerbach, P.P. (1990) the influence of prior knowledge and text genre on readers' prediction strategies. *Journal of Reading Behavior* (Chicago, IL), vol. 22, no. 2, p. 131–48.

2. Alderson, J. C., & Urquhart, A. H. (1983). Reading in a Foreign Language. London, New York: Longman.

3. Alderson, J. C. (1996). The Testing of Reading. In Nuttall, C. (ed.) *Teaching Reading Skills in a Foreign Language* (2nd Ed.) Oxford: Heinemann English Language Teaching, 221-29.

4. Alderson, J. C. (2003). *Reading and use of English*. Budapest: Teleki Laszlo Foundation.

5. Alderson, J. C., Figueras, N., Nold, G., Kuiper, H., Takala, S. e Tardieu, C. (2004). *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: learning, teaching, assessment.* Final Report of the Dutch CEF Construct Project. Unpublished manuscript.

6. Alderson J. C., (2007), Judging the Frequency of English Words, *Applied Linguistics* 28/3, 383–409.

7. Alekseev, P. M. (1984). *Statistische Lexikographie*. Bochurn: Brockmeyer.

8. Anderson, R.C. & Freebody, P. (1981). Vocabulary Knowledge. In J. Guthrie (ed) *Comprehension and Teaching: Research Reviews*, Newark, DE: International Reading Association, 77 – 110.

9. Anderson, R.C. & Pearson, P.D. (1984) A schema. Theoretic view of basic processes in reading comprehension. In P.D. Pearson (ed), *Handbook of Reading research*, New York: Longman. 255-291.

10. Aronoff, M. (1994), *Morphology by Itself*. Cambridge, MA, Mit Press.

11. Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

12. Bachman, L F (2004). *Statistical Analysis for Language Assessment*. Cambridge: Cambridge University Press.

13. Balboni, P. (2000). *Le microlingue scientifico –professionali: natura e insegnamento*. Torino: Utet.

14. Baldwin, R.S. Peleg-Bruckner, Z. & McClintock, A.H. (1985) Effects of topic interest and prior knowledge on reading comprehension *Research Quarterly* 20, 497 -504.

15. Beccaria, G. L. (ed.) (1996). *Dizionario di linguistica e di filologia, metrica, retorica.* Torino: Einaudi.

16. Bernhardt, E. B. (1984) Toward an information processing perspective in foreign language reading. *Modern Language Journal*, 68, 322 – 331.

17. Bianco A*, (*2002) *La linguistica dei corpora*, Lecce: Adriatica Salentina.

18. Biber, D., Conrad, S. e Reppen, R. (1998). *Corpus linguistics: Exploring language structure and use.* Cambridge: Cambridge University Press.

19. Borello, E. (1994). *L'incomunicabilità di massa – Linguaggi settoriali: funzionamento e apprendimento.* Alessandria: Edizioni dell'Orso.

20. Bowker, L. & Bennison, P. (2003). Student translation archive: design, development and application. In Zanettin, F., Bernardini, S. e Stewart, D. (eds.) *Corpora in translator education*. Manchester: St Jerome, pp. 103-17.

21. Bugel, K. & Buunk, B. (1996) Sex differences in foreign Language text Comprehension: the role of Interests and Prior Knowledge, *The Modern Language Journal* 80/1, 15 -31.

22. Bureau for International Language Coordination (BILC) (2002): Testing seminar material, Version 2.1.

23. Carrell, P.L. & Wise, T. E. (1998) Prior Knowledge and Topic Interest in Second Language Reading, SSLA, 20, 285 – 309.

24. Chiari I., (2007) *Introduzione alla linguistica computazionale,* Roma: Laterza.

25. Chung, T. M. & Nation, P. (2003) Technical vocabulary in specialized texts, *Reading in a Foreign Language* 15/2, 103 -116.

26. Clapham, C. (1996) *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

27. Davidson, F. (1998) Book Review: The development of ILTS. A study of the effect of background knowledge on reading comprehension. *Language Testing* 15, 288-292.

28. De Mauro, T. (2003). *Grande dizionario italiano dell'uso*. Torino: UTET.

29. Douglas, D. (2000) *Assessing Languages for Specific Purposes*. (eds. Alderson, J.C. & Bachman, L.F.) Cambridge: Cambridge University Press.

30. Douglas D. (2001), Language for Specific Purposes assessment criteria: where do they come from? *Language Testing* 18/2 171–185.

31. Garside, R., Leech, G. & Mc Enery, A. (1997). *Corpus Annotation, Linguistic Information from Computer Text Corpora*. London: Longman.

32. Geeraerts, D. (1984). Dictionary classification and the foundations of lexicography. *ITL Review* 63, 37-63.

33. Gigliozzi G. (2003*), Introduzione all'uso del computer negli usi letterari*, a cura di F. Ciotti, Mondadori, Milano.

34. Gotti, M. (1991). *I linguaggi specialistici: caratteristiche linguistiche e criteri pragmatici*. Firenze: La Nuova Italia.

35. Granger, S. (ed.) (1988). *Learner English on computer*. London & New York: Longman.

36. Green, B., Wolf, A., Chomsky, C. & Laugherty, K. (1961). *BASEBALL: An automatic question answerer.* Proceedings of the Western Joint Computer Conference (19), 219-24.

37. Green, R. & Wall, D. (2005) Language Testing in the Military: Problems, Politics and Progress, *Language testing* 22/3 379 – 398.

38. Habert, B., Nazarenko, A. & Salem, A. (1997) *Les linguistiques de corpus*. Paris: Armand Colin/Masson.

39. Hammadou, J. (1991) Interrelationships among Prior Knowledge, Inference and Language Proficiency in foreign Language Reading, *The Modern Language Journal* 75/1 27- 38.

40. Hamp-Lyons L. & Lumley T., (2001) Assessing language for specific purposes, *Language Testing 18*, 127-132.

41. Jennings, M. Fox, J. Graves, B & Shohamy, E. (1999) The effect of choice on test takers' scores in language tests. *Language Testing*. 15, 426- 456.

42. Johnson, P. (1982) Effects of reading comprehension on building background knowledge. *TESOL Quarterly*, 16, 503-516.

43. Laufer, B. & D.D. Sim (1985) Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals.* 18, 405 - 411.

44. Leech, G., Garside. R. & Bryant, M. (1994). The large-scale grammatical tagging of text: experience with the British National Corpus. In Oostdijk. N. e de Haan, P. (eds.) *Corpus-based research into language*. Amsterdam: Rodopi, 47-63.

45. Lynch, Bk. & Davidson, F. (1997). Criterion referenced testing. In Clapham, C. e Corson, D. (Eds.) *Encyclopaedia of Language and Education: Language Testing and Assessment* vol. 7. Dordrect, The Netherlands: Kluwer Academic Publishers 263-273.

46. McEnery, T. & Wilson A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University.

47. Muller, C. (1979). Sur les emplois personnels de l'indéfini. *Langue française et linguistique quantitative.* Genève: Slatkine 65-72.

48. Papageorgiu, A. (2007). *Teaching English for Special Purposes to military officers: utopia or reality within the classroom environment*? In www.besig.org/events/conf2007/downloads/Papageorgiou3.doc. (Last accessed 1 September 2008).

49. Papajohn, D. (1999) The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants. *Language Testing* 16/52. 52-79.

50. Pearson, J. (2003). Using parallel texts in the translator training environment. In Zanettin, F., S. Bernardini e D. Stewart (eds.) *Corpora in translator education*. Manchester: St Jerome 15-24.

51. Peretz, A.S. & Shoham, M. (1990). Testing reading comprehension in LSP.; Does topic familiarity affect assessed difficulty and actual performance? *Reading in a Foreign language*, 7. 447 – 455.

52. Pike, L. W. (1979). An Evaluation of Alternative Item formats for Testing English as a Foreign Language*. TOEFL Research Reports 2*. Princeton (NJ): Educational Testing Service.

53. Read, J. (2000) *Assessing Vocabulary* (eds Alderson, J.C. & Bachman, L-F.) Cambridge: Cambridge University Press.

54. Rossini Favretti, R. (2000). Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. In Rossini Favretti, R. (ed.) *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Roma: Bulzoni, 39-56.

55. Schmitt N., (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge, *Language Testing* 16/2, 189–216.

56. Shohamy, E. (2001). *The power of tests: a critical perspective on the uses of language tests.* Harlow: Longman.

57. Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

58. Sinclair, J. (1992). *Collins COBUILD English language dictionary.* London: Glasgow.

59. Sinclair, J. (1996). The search for units of meaning. *Textus* 9/1, 75-106.

60. Spilich, G.J. Vesonder, G.T., Chiesi, H.L. & Voss, J.F (1979) Text Processing of domain-related information for individuals with high and low domain knowledge. *Journal of verbal Learning and verbal Behavior*, 18, 275-290.

61. Spolsky, B. (1995). *Measured Words.* Oxford: Oxford University Press.

62. Stahl, S.A, Jacobson, M. G. Davis, C. E., & Davis, R.L. (1989) Prior knowledge and difficult vocabulary in the comprehension of unfamiliar Text. *Reading Research Quarterly*, 14., 27-43.

63. Standardized Agreement (STANAG) 6001 NTG (EDITION 1) (1976): Language Proficiency Levels.

64. Standardized Agreement (STANAG) 6001 NTG (EDITION 2) - LANGUAGE PROFICIENCY LEVELS http://www.bilc.forces.gc.ca/stanag/index-eng.asp, (Last accessed July 6 2009)

65. Tan, S.H. (1990) The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In J. H.A.L. de Jong & D.K. Stevenson (eds). *Individualizing the assessment of language abilities.* Clevedon, England: Multilingual matters: 214-224.

66. Taylor, L. Thompson, P. Mc Carthy. & Barker, F. (2003), ' *Exploring the relationship between language corpora and language testing'*, Symposium at 25th Language Testing Research Colloquium, Reading, UK, 22-25 July 2003.

67. Taylor L. & Barker, F. (2008) '*Using Corpora for Language Assessment'* in E. Shohamy & N.H. Hornberger, N.H. (eds) Encyclopedia of Language and Education, 2nd edition, Volume 7: Language Testing and Assessment, 241- 254. Springer Science Business Media LLC.

68.   Vietri, S. & Elia. A. (1999). *Electronic Dictionaries and Linguistic Analysis of Italian Large Corpora.* ACL Anthology. March 24 1999. http://www.aclweb.org/anthology-new/W/W99/W99-0513.pdf. (last accessed 25/08/08).

69.   Winograd, T. (1983). *Language as a Cognitive Process. Vol. I: Syntax.* Reading (MA): Addison-Wesley.

70.   Zanettin, F. (1998). Bilingual Comparable Corpora and the Training of Translators. *Meta* 43/4, 616-30.

LIST OF APPENDICES

Appendix 1 STANAG 6001 Edition 1 (1976) – excerpt of reading comprehension level 3 proficiency

NATO Standardization agreement (STANAG)

Language proficiency levels STANAG 6001 - Navy/Army/Air Force

1. Aim	4. Proficiency levels
The aim of this agreement is to provide the NATO Forces with a table describing language proficiency levels.	The proficiency skills are broken down into five levels coded 1 through 5. In general terms, skills may be defined as follows:
2. Agreement Participating nations agree to adopt the table of language proficiency levels for the purpose of:: a. Meeting language requirements for international staff appointments. b. Comparing national standards through a standardized table. c. Recording and reporting, in international correspondence, measures of language	Level 1 Basic Level 2 Functional Level 3 Professional Level 4 Expert Level 5 Excellent (Native/bilingual) Code 0 indicates there is no significant or practical proficiency in the skill concerned.

proficiency (if necessary by conversion from national standards).	
3. General	
The descriptions at page 3 and 4 give detailed definitions of the proficiency levels in the commonly recognized language skills: oral proficiency (listening and speaking) and written proficiency (reading and writing).	

Level 3- Professional Adequate for standard text materials and most technical material in a known professional field; with moderate use of dictionary, adequate for most news items about social, political, economic and military matters. Information is obtained from written material without translation.

Appendix 2   STANAG 6001 Edition 2 (2001) – excerpt of reading comprehension level 3 proficiency

LEVEL 3 (MINIMUM PROFESSIONAL)

READING COMPREHENSION

Able to read with almost complete comprehension a variety of authentic written material on general and professional subjects, including unfamiliar subject matter. Demonstrates the ability to learn through reading. Comprehension is not dependent on subject matter. Contexts include news, informational and editorial items in major periodicals intended for educated native readers, personal and professional correspondence, reports, and material in special fields of competence. Can readily understand such language functions as hypothesizing, supporting opinion, argumentation, clarification, and various forms of elaboration. Demonstrates understanding of abstract concepts in texts on complex topics (which may include economics, culture, science, technology), as well as his/her professional field. Almost always able to interpret material correctly, to relate ideas, and to "read between the lines," or understand implicit information. Can generally distinguish between different stylistic levels and often recognizes humour, emotional overtones, and subtleties of written language. Misreading is rare. Can get the gist of higher level, sophisticated texts, but may be unable to detect all nuances. Cannot always thoroughly comprehend texts that have an unusually complex structure, low frequency idioms, or a high degree of cultural knowledge embedded in the language. Reading speed may be somewhat lower than that of a native reader.

Appendix 3   TUI specifications

1. The T.U.I has been changed to adhere faithfully to the STANAG 6001 Amplified. This agreement defines a candidate's levels in a particular language and the standards and range within each level.

2. The T.U.I. consists of four individual tests:

   o a **listening** comprehension test

   o a **speaking** test

   o a **reading** comprehension test and

   o a **writing** test.

1. The level received in each one of these tests becomes the candidate's **SLP,** which is **Standard Language Profile.** An **SLP** of **3332** means a level **3** in Listening, a level **3** in Speaking, a level **3** in Reading, and a level **2** in Writing.

The **reading comprehension** test consists of 60 multiple choice questions which have been broken down into levels as in the listening test:

level 1	18 questions
level 2	17 questions
level 3	13 questions
level 4	12 questions

The questions have been grouped into sections according to the type of reading activity.

news items	as per STANAG 6001 Amplified level 3	5 questions
articles & extracts	as per STANAG 6001 Amplified level 3	4 questions
military and professional texts	as per STANAG 6001 Amplified level 3	4 questions

N.B. All items and passages have been taken from authentic texts.

EXAMPLES of some of the questions in the READING COMPREHENSION test

NEWS ITEMS (level 3)

Yesterday we reported that the Foreign Office had advised tourists not to visit Luxor. The Foreign Office advises tourists not to travel by road, rail or river to or through the Egyptian governorate of Minya. This is because of the number of terrorist attacks against foreign visitors to rural Egypt in the last few months.

Tourists to Egypt...

> a. *reported to the Foreign Office because of the terrorist attacks.*
> b. *received no warnings from the Office while touring Minya.*
> c. **should avoid the Minya region at least for the time being.**
> d. *must visit Luxor before they leave for the tour to rural Egypt.*

READING EXAMPLES continued...

ARTICLES & EXTRACTS

(level 3)

A global smoking ban for airline passengers moved a step closer yesterday when British Airways announced the extension of a non-smoking policy to almost all its flights. In response to customer pressure, from next year, an extra 350 flights a day to Africa, Europe, the Middle East, the Far East, and North America will be designated non-smoking for a trial period, in addition to 750 existing smoke -free services.

The ban will cover all BA's transatlantic flights for the first time, including the Concorde service to New York costing £5,000 for a return ticket. Previously, 12 of the supersonic aircraft's 100 seats were for smokers. Travellers still wishing to smoke on a UK - US flight will have to switch from BA to one of the dwindling number of airlines on which it is still permitted. Virgin Atlantic banned smoking on flights to America last year.

According to this article...

> a. *smoking is forbidden on all BA flights.*
> b. *all London to New York flights will be no-smoking.*
> c. **BA will introduce an innovation in 350 new flights.**
> d. *there are only 12 places for smokers on flights to the United States.*

MILITARY & PROF TEXTS   **(level 3)**

Buck Knives Co., El Cajon, California. USA

$10 million

From: US Army

For: approximately 250,000 M9 bayonets for the US Army. Beginning this year, the contract calls for optional quantity deliveries over its five-year span. The M9 is a rugged field knife with a «zone heat-treated» 18 cm forged-steel blade. The blade has a saw-toothed back edge that will cut rope, ice, and most aircraft fuselage metal. When coupled with a stud on the rugged plastic sheath it becomes a wire cutter.

Buck Knives. Co...

   a. *is used to dealing with bayonets from the US Army.*
   b. *has spent $10 million developing a new knife.*
   c. ***will produce weapons for the American army.***
   d. *has delivered 250,000 knives to the US armed forces this year.*

Appendix 4   JFLT specifications

Reading Comprehension

JFLT Test structure:

Techniques: Examinees read a text and answer multiple-choice questions

Medium: paper and pencil, computer

Test time: 105 minutes

Format

15 authentic Level 3 texts, with comprehension evaluated by 4-option multiple-choice items.

Instructions will be in the target language

Criterial levels of performance:

A pre-established number of items must be responded to correctly to fulfil level requirements.

Content

Operations or Functions

Operations are based on tasks and accuracy requirements found in the STANAG 2^ ed. descriptors for Level 1 to 4.

Level 3

Tasks:

Obtain gist of higher level and sophisticated texts (but cannot detect all nuances)

Hypothesizing

Supporting opinion

Argumentation

Clarification

Various forms of elaboration

understand implicit information ("read between the lines")

Can generally distinguish between different stylistic levels

Can often recognize humour, emotional overtones, and subtleties of written language

Content:

Authentic written material

News, informational and editorial items in major periodicals intended for educated native readers

Personal and professional correspondence

Reports

Material in special fields of competence

Abstract concepts in texts on complex topics

Topics:

General and professional subjects, including unfamiliar subject matter

Complex topics which may include economics, culture, science, technology

His/her professional field

Accuracy & Language Conditions:

Demonstrates the ability to learn through reading

Comprehension is not dependent on subject matter

Misreading is rare

Cannot always thoroughly comprehend texts that have an unusually complex structure, low frequency idioms, or a high degree of cultural knowledge embedded in the language

Reading speed may be somewhat slower than that of a native reader

JFLT SPECIFIC

Types of text/task:

Texts from authentic sources written by native speakers for native speakers and not intended for instructional purpose (e.g., academic texts taken from textbooks, journal articles).

Texts included in the test are to be representative of the texts candidates should be able to read successfully (for content validity and "backwash"; a test of only a limited range of texts will encourage the reading of a narrow range of texts by potential candidates).

Level 3 (number of items / type of text)

5   materials in special fields of competence, professional subjects

3   general topic reports

1   personal & professional correspondence

6   excerpts / articles (e.g. academic/informational, periodicals, news articles, editorial items,

Appendix 5   Transcription of interview - TUI test

FG   Ok, after this short practice with the objects in the room we can begin.

AS   (number 1) I went too fast. I better read it one more time….

FG   Where do you think this text was taken from?

AS   From a novel.

FG   Ok. Are you familiar with military fiction?

AS   A little bit, yes

   ….so, here the factory is in a small valley…. And so….carefully hidden because it was invisible from the air…

FG   (number 2) How do you like this?

AS   This is tricky. Because …. (reading and scanning the text line by line – reading more than once and getting a bit uncomfortable. He excluded three distracters and chose A).

FG   Why did you find it tricky?

AS   Because the distracters have the reverse meaning, so you've got to pay special attention

FG   All right, let's move to the next (number 3)

AS   (scanning very quickly and going straight to the correct answer)

FG   Ok, so you went very straight on this one...

AS   yes

FG   No problem at all.

AS   yes

FG   Very clear on the definition of Armed reconnaissance

AS   yes

FG   Ok, good, next (number 4)

AS   (scanning the text more than once) this is D

FG   Ok, why?

AS   Because….. (pointing out and reading the very line and excluding all the others distracters

FG   (number 5)

AS   This is C, no doubts. See? It does not say directly but, it is clear that you've got to avoid possible local strives

FG   All right. So, Next (number 6)

AS   (reading out loud the text). I think D, I am not sure but this is better than the other choices

FG   (number 7)

AS   (reading out loud the text, then the distracters) this is really confusing, I could choose all of them and yet….

FG   What makes you think this?

AS   (reading the text one more time – he seems lost and unfocussed, as if he would just get rid of this question) I think this is B

FG   All right. Where have you found the catch to answer this question?

AS   It is not clearly said in the text, so by excluding the others options.

FG   Ok, good! Next (number 8)

AS   No doubts. This is D

FG   Why? How did you get this?

AS   The others seem inconsistent.

FG   So, it seems like you could answer this without even reading the text. Isn't that?

AS   Yeah!

FG   Well, let move on then (number 9)

AS   (After a very long scan) thirteen is B

FG   What skill do you think you needed to answer this one? Did you find in the text what you were looking or ….

AS   I deducted it from the content

FG   Ok, what about number ten?

AS   fourteen is B because in the text it is stated what the use of the attached list should be. No doubt about that.

FG   Ok. Good. Do you have any other comment on this test? How did you find it?

AS   It thought it was quite difficult. The enemy is the time, you know. You've got to read, read again.

FG   Did you have to read between the lines?

AS   Yes. More than once.

FG   What is your feeling about this military-flavoured kind of test?

AS   Very useful.

FG   You do prefer to deal with military stuff!

AS   Yes. This is the kind of stuff I work with and it sounds more familiar. I prefer to deal with this rather than more general topics. At least you know the atmosphere. It can be dangerous though, because you think to be in safe waters….

FG   Ok. Well then. Thank you for your help.

Appendix 6   Transcription of interview - JFLT test

FG   All right. Here you have 10 questions. You will read through them and answer the question at the bottom of each one. Remember, I am interested mostly in finding out what are the strategies you use to answer your questions.

AS   Ok. ……. (quite often he reads chunks of text out loud) …. This one is tricky. This is C

FG   Ok. Why so?

AS   There is an implied meaning. In the first part of the paragraph they talk about water as vital source, transboundary water issues, and etcetera. So, it cannot be A. B is partially right, and last one…. Cannot be. So my choice is C

FG   Did you read the distracters first and then you went on the text, or just the contrary?

AS   I begun from the text

FG   Is this your usual technique?

AS   Yes

FG   Good. Next (number 2)

AS   My choice is B; prison is not a good response to juvenile delinquency. It cannot be neither A, C or D.

FG   Still, it is not clearly stated. So, you deducted it from the text, isn't that?

AS   Yes.

FG   Ok. Good. Next. (Number 3).

AS   I chose B. according to the text it's the only one that fits the

description, (then he starts reading it again)…. Mmmhh A could also be right. I go for B anyway

FG   So, you scanned the text and then excluded the wrong ones.

AS   Yes. I checked twice to be honest, because it was a bit confusing. It could be A or B.

FG   Ok. Next then (number 4)

AS   Number 4 is D. The presence of fishes in rice fields. It is not A.

FG   Although fish are never mentioned….

AS   Well, indirectly yes. They are

FG   Good. Next (number 5)

AS   This is B. very easy. Only two lines of description. Very factual.

FG   Good. Next.

AS   This is A. Very easy. An eye for an eye.

FG   Good. Next. (number 6)

AS   Now I am going to change strategy. I will read the answers first and then the text.

FG   Ok.

AS   I exclude A. as for B, mmmmh I don't know. I am reading again the answers…. (reading out loud the answers, more than once) … this is B

FG   So, this time you had to read it again, uh?

AS   Yes. For me the possible answers are A or B. as for C …. No, it is not clearly expressed, although he (the author) mention the invasion in Granada and then, is critical toward another (military) operation

FG   Where does it say that it does not support?

AS   It is an implied meaning, not clearly said.

FG   Still, in the first part of the article, the author is talking about language. Right?

AS   Yes, but…

FG   … and the use of language in order to…

AS   …words that are used to conceal rather than reveal meanings…

FG   …maybe that could be the key to decode the whole context

AS   I see, so maybe D is much more appropriate

FG   All right. Next (number 7)

AS   This is A

FG   These are difficult, uh?

AS   Right. You have to read them, and then make your own evaluation… your deductions by yourself. I had to read it twice

FG   Ok. Next (number 8)

AS   This one is B. None of the others can fit.

FG   Ok. Next. Are you still reading the answers first?

AS   Yes, I prefer in that way.

FG   Ok.

AS   A by listing a set of criteria

FG   Yeah, there are so many examples, right? Ok. Next (number 9)

AS   I go for C. Wait, let me read again just to be sure. Mmmhhh I go for D.

FG   So, what is your evaluation about this test?

AS   Well, this one covered many different issue, not military though. Honestly, it is not very important that they are only military. Some of the passages needed more checks (care) in order to catch the best answer, and sometimes I had strong doubts about the right answer. In some case I had to re-read the entire text, but this is part of the game.

FG   Yes, of course. One more question. Would you prefer a test of military flavour or do you prefer a more general and wider span?

AS   It could be argued that a test covering more issue other than military might give me more chances… but I think this is not the real issue. Time constraint is the real thing, that is trying to manage and balance the attention needed within the time allotted

FG   Well, Alessandro. Thank you very much for your time

Appendix 7   TUI selected items for administration[7]

> The plane was flying behind enemy lines in the early hours of the morning when the parachutists jumped, with instructions to gain as much information as possible on a new experimental factory. They had been told that the building was in a small valley and so carefully hidden that it was invisible from the air. Both men knew their task was not dangerous - the roads to the factory were well guarded but there were not many troops stationed in the area.
>
> The men had destroyed their parachutes and had dressed as workers. They went across some fields to a road where they thought the factory was located. Along the road they saw a lorry parked beside a deep hole. They were cautiously walking towards the lorry when they heard the sound of a car in the distance. They ran for cover, but there was so little vegetation in the valley that they had no alternative other than to jump into the hole. They listened in silence as the car came nearer, hoping no-one had seen them. The car stopped almost directly above them and some men got out. The parachutists heard voices and, from the conversation that followed, understood the men had come to collect the lorry. They did not move or make a sound until the vehicles had left. Then they both smiled at the thought of not having been captured.

1.   The parachutists were not sure of the location of the factory because...

       a. it was invisible.
       b. it was well-hidden.
       c. it was in a valley.
       d. it was well-guarded.

---

[7] The items reproduced in this appendix were included in the now de-classified TUI test. Reproduction of the ten items from the JFLT used in the study is forbidden under military regulations as the test is still classified material.

Friendly targets are represented by one of the following symbols, as appropriate. Targets for friendly fire are normally designated using two letters followed by four numbers. For additional information on these symbols, see FM 6 - 20.

2. The symbols...

a. are designated by friendly targets.
b. are represented by friendly targets.
c. are not explained in full detail.
d. are appropriate to the numbers given.

ARMED RECONNAISSANCE - A mission with the primary purpose of locating and attacking targets of opportunity (such as, enemy materiel, personnel, and facilities) in assigned general areas or along assigned ground communications routes, and not for the purpose of attacking specific targets.

3. This definition explains that...

a. any target may be located but not attacked.
b. the targets don't have a specific purpose.
c. the targets can't be attacked unless specified.
d. any target may be located and attacked.

Continuum of Military Operations

Future Army activities will be conducted throughout the continuum of military operations. The strategic environment within each theatre consists of a variety of conditions - political, economic, military - and a range of threats that result in varied responses. The continuum of military operations is an analytical construct which links the strategic environment and threats within a theatre to appropriate military actions. This continuum consists of three general states: peacetime engagement, hostilities short of war, and war.

Peacetime engagement represents a predominantly non-hostile state of the continuum characterized by the benign use of military forces along with political, economic, and informational measures to achieve national objectives and to complement our efforts to deter conflict or, should deterrence fail, win in combat. Operations in support of peacetime engagement are normally interagency in character and are designed to address the fundamental causes of instability that can lead to regional conflict.

Hostilities short of war involve armed struggle between organized parties within a nation or between nations in order to achieve limited political or military objectives. While conventional forces may be involved, special operations forces or non-combat forces frequently predominate. Hostilities short of war are often protracted, confined to a restricted geographic area and constrained in weaponry and level of violence. Limited objectives may be achieved by the short, focused, and direct application of force.

War involves the sustained use of armed force between nations or organized groups within a nation employing appropriate military forces in a series of connected battles, major operations and campaigns to achieve vital national objectives. War may be limited or it may be general, with the total resources of a nation employed and the national survival of a belligerent at stake

According to this text:

4.   **the strategic environment...**

a. solely involves political, economic and military conditions.

b. is built up through an analysis of operations.

c. is the consequence of a variety of responses.

d. is linked to appropriate military actions.

5.   **peacetime engagement...**

a.   supports mainly interagency operations.

b.   is the continuum without any form of hostility.

c.   involves strategies to avoid local strife.

d.   excludes any possible military employment.

6.   **war...**

a.   is exclusively between countries.

b.   requires that all the natural resources of a country be used up.

c.   implies a continuous employment of armed force.

d.   encompasses all neighbouring countries.

The Army is guided by four enabling strategies to meet the challenge of the future

To **Maintain the Edge** in war-fighting that was demonstrated in Panama and the Persian Gulf by balancing the imperatives that ensure our success - quality soldiers who have been trained to razor sharpness, supported by dedicated civilian employees, outfitted with the most modern equipment, led by tough and competent leaders, structured in an appropriate mix of forces by component and type, and employed according to an effective war-fighting doctrine.

To **Reshape the Force** by tailoring active and reserve forces to accommodate the new strategy. We are reducing our presence in Europe by nearly 60%; we have already eliminated two divisions in COONS; and we must be allowed to reduce the Cold War vestige of unneeded reserve component force structure - and return to programmed reductions this fiscal year. The Army will also recast training and war plans to accommodate a regional focus and the rapid deployment capability of units based in the continental United States, and evolve our doctrine to reflect changes in modern battlefield dynamics as well as the emphasis on joint, combined, and coalition warfare and the use of Army forces across the continuum of military operations.

To **Provide Resources for the Force** by improving the force structure to preserve readiness despite budget constraints, by making tough management decisions, and by becoming ever more efficient stewards of America's treasure.

To **Strengthen the Total Force** by fully integrating our active and reserve components, keeping early-deploying units fully 'mission ready' establishing strong training relationships and by fully integrating Total Army readiness standards and operating systems.

The Army is already changing to reach the force structure required to execute the National Military Strategy. The Base Force of 1995 is one-third smaller, and it represents the right combination and mix of forces and capabilities to deter potential aggressors, fulfil our forward presence requirements and respond to any crises which threaten vital US interests. As the Army reduces its size, we must sustain our investments in leader development and quality soldiers, continue to evolve our very sophisticated training programs, and bring new and more discriminating training strategies to the Army Guard and Reserve. The forces that we maintain must be trained, sized and equipped to ensure they

possess an edge in effectiveness that no potential adversary can match. They must be capable of winning the first battle.

7.   To meet the challenge of the future, the Army...

   a. will be employing professionals to see to their budg*et al*lotment.
   b. needs to focus more on possible local strife than ever before.
   c. must show foreign nations how successful it was in previous conflicts.
   d. has to be better trained in peace-enforcing missions.

8.   The Army can restructure the force...

   a. by accommodating personnel in the best manner.
   b. by best accommodating the national military strategy.
   c. by not placing too much emphasis on joint and coalition warfare.
   d. by adhering more closely to current strategies in their doctrine.

SECTION 5   CORRESPONDENCE

NATO UNCLASSIFIED

Headquarters

ACE Rapid Reaction Corps

41179 Edilberg
Highland Forces

Trinton   **Post Office 40**

Telephone:

JHQ Edilberg Extension 12374

Telecom 055 - 37 - 321972

Fax 43520

2331.17/BSDTFM/97   17
December 1997

TO: See distribution

SUBJECT: CORPS ENGINEER TRAINING

FUNCTIONAL PRIORITY LIST

1.   Corps engineering comprises a very wide spread of technical skills and collective training demands. In order to assist Commanders of ARRC assigned engineers to focus their training in collective skills and on individual experts, I wish to define groups of tasks at which to aim training.

2.   To this end, would you examine the attached list and indicate to me firstly, those for which you are trained or have expertise, regular or reserve, now, and secondly, those skills or groups of skills which you would be prepared to take on in the future. I believe a battalion-sized unit should be prepared to put the emphasis on two of the areas on the list, with some overlap to other units but not wholesale duplication. If you have identifiable sub units or small groups or agencies (e.g. Highland Specialist Teams) it would be helpful if you could describe the levels of expertise and experience they contain.

3.   Your replies, with the agreement of your national authorities are

required by 3 Feb. 1998.

SEAN J GLENGRANT

Brigadier Highland A

1

NATO UNCLASSIFIED

9.   This letter states that the General...

a. is informing his commanders as to how to proceed with the engineers' training.
b. means to explain the skills, abilities and training which are needed by a corps engineer.
c. wishes to focus on individual experts rather than on wide-spread collective abilities.
d. is concerned with engineers' individual expertise rather than in individualized training.

10.   The attached list...

a. is needed to point out the engineers' training ability to their commanders.
b. needs to be checked by each commander as to his/her present and future expertise.
c. itemizes the commanders' prior abilities and expertise.
d. shows where the corps engineers can train future engineers in their battalions.