UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

DIPARTIMENTO DI SCIENZE POLITICHE
SCUOLA DI DOTTORATO IN
SCIENZE PSICOLOGICHE, PEDAGOGICHE E LINGUISTICHE

DOTTORATO DI RICERCA IN
LINGUA INGLESE PER SCOPI SPECIALI
XXIV CICLO

TESI DI DOTTORATO

DISSEMINATING STATISTICS IN EUROPE: THE LANGUAGE PERSPECTIVE

CANDIDATA
DOTT.SSA FRANCESCA SCAMBIA

RELATORE
PROF.SSA VANDA POLESE

COORDINATORE
PROF.SSA GABRIELLA DI MARTINO

NAPOLI 2013

CONTENTS

ACKNOWLEDGMENTS

I would like to express my thanks to all those who in many ways have contributed to enrich this study.

I wish to thank all the members of the Ph.D Board and in a special way the coordinator, Professor Gabriella Di Martino, for her encouragement and support.

I am very grateful to my supervisor, Professor Vanda Polese, for guiding me with constructive criticism and suggestions and assisting me with care in revising my Ph.D work.

I also wish to thank Professor Silvia Bernardini who has believed in this project and has helped me with her knowledge and openness.

I owe many thanks to my PhD colleagues for being friendly and cooperative in exchanging ideas and building an enjoyable environment while sharing such a stimulating and enriching experience.

I am also grateful to my ISTAT colleagues who have facilitated the completion of this work providing suggestions and useful information.

Needless to say, any inaccuracies in the research are entirely my own responsibility.

INTRODUCTION

Statistics is a science dealing with numbers. At the entrance of the ISTAT (Italian National Statistical Institute) headquarters in Rome a Latin motto is carved to remind people of the importance of numbers – *Numerus omnium rerum nodus. Numerus respublicae fundamentum*[1]. Numbers have wide applications. That is why in the last decades an ever-growing use of statistical data in decision-making fields has caused statistics to play an active and major role in the management of public life. Nowadays, social, political, economic speeches and hardly any text very often quote and interpret statistics. From such widespread use of data also stemmed the need to make statistics available to a wider public. One of the ways to reach out the international community is by means of English as the main communication language. And indeed at European level publications on national statistics have begun to be more and more frequently translated into English in the last decades.

In her study "Making Sense of Statistics"[2], Jessica Gardner (2009), from the United Nations Economic Commission for Europe (UNECE), remarks that:

> Effective presentation of data is increasingly being considered an integral part of the statistical production process. Clear and simple presentations of statistics are needed to ensure they are widely used and correctly interpreted. As user groups diversify, many statistical organizations are reassessing their approach to presentation to take advantage of new technology and address changes in user needs.

This diversification of user groups is indeed changing the way statistics' knowledge is being disseminated. The present study focuses on the

[1] Number is the crux of everything. Number is the foundation of Republic (my translation).

statistics presented in the English language and disseminated within the EU.

Statistical data are produced by several private, national and supranational agencies. Not all statistical output can be considered official and only some specific agencies are recognized by the law, at national and European level, as producers of official and hence trustworthy statistics.

Referring to official statistics, on 20 October 2010, the United Nations promoted the first World Statistics Day. This is further evidence that statistics is spreading over the years and its social role is growing more and more important in determining, for instance, the existing position of *per capita* income, unemployment rates, population growth rate, housing, schooling, medical facilities, etc. in a country. Furthermore, it holds a central position in almost every field of human activity like industry, commerce, trade, physics, chemistry, economics, mathematics, biology, botany, psychology, astronomy, etc.

Statistics is thus essential for a country. Different policies of the government are based on statistics and statistical data are now widely used in taking all administrative decisions. If a government, for instance, wanted to revise pay scales for employees, statistical methods would undoubtedly be used to determine the rise in the cost of living. The preparation of government budgets mainly depends upon statistics because statistical data help estimate the expected expenditures and revenue from different sources. Thus, statistics is the eyes of a state administration. This is just an example which proves that the increasing relevance of statistics is recognised worldwide. Nevertheless, very little

[2] http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/1242.pdf (Last accessed 10 December 2012)

has so far been done to study and analyze the written language of statistics for efficient and effective communication. Language in statistics has hardly ever been considered particularly relevant, since statisticians can interpret tables and graphs with no need for further explanations, just like mathematicians read formulas. However, the interest shown by a larger public has impelled statisticians to comment on and provide explanations to the statistical output in a way that is understandable by laypeople as well.

The purpose of this research is to investigate official statistics within the European Union, that is, the statistical output by member-states and Eurostat (the EU statistical bureau). The main research questions concern whether, how and to what extent the principles of clarity and accessibility have been considered in the presentation of statistical data, and what are the main differences between national and EU publications from the language viewpoint. Such questions involve the issue of clear writing which has been a concern of the EU in its effort to reach out EU citizens and act in accordance with the European "Fight the Fog"[3] campaign. Specifically, the research thus aims to investigate differences between statistical publications by the EU and texts drafted by the National Statistical Institutes. It also arises questions on the use of translation to make national statistical texts available to an international public. Since the English translation of national statistics is addressed to an international and intercultural public, it can be included in the field of intercultural communication (Kankaanranta 2009)[4]. As a whole, the study

[3] http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/issue_01_en.pdf (Last accessed 10 November 2012)

[4] http://www.languageatwork.eu/downloads/LAW%20Business_English_Lingua_Franca_in_intercultural_business_communication.pdf (Last accessed 2 January 2013)

of statistics' language is pretty new and therefore needs a great deal of effort and investigation. This study aims to be a contribution to the development of this branch of ESP.

1. THEORETICAL FRAMEWORK

In this chapter I shall describe the EU statistical framework and regulations with respect to the new role of statistics and presentation of statistical knowledge and data. Specifically, the English language of statistics will be discussed with a focus on statistics as specialized discourse, the use of English as a Lingua Franca (ELF) and the English translation of statistical publications. All these aspects will be related to the goal of intercultural communication within the European context.

1.1 The European Statistical System

The European Statistical System (ESS) is based on the National Statistical Institutes usually referred to as NSIs. All NSIs of EU member countries are required to produce official statistics and provide data to Eurostat. Eurostat is the statistical office of the European Union located in Luxembourg. Its task is to provide the EU with statistics at European level in such a way as to enable comparisons between countries and regions "in view of a common statistical language", as we can read on the Eurostat website[5].

The challenge that Eurostat has to face concerns the building of communication channels among different European languages and cultures. As for statistics, the establishment of a 'common language' also involves the use of the same parameters and a common way of drafting tables and graphs, although a further concern is the increased demand for

[5] Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/about_eurostat/corporate/introduction (Last accessed 18 October 2011)

statistical information by non-specialized users which has led statisticians to increase the verbal elements in written statistical texts. Words are indeed the most understandable way of communicating out of a specific scientific community. This phenomenon is particularly evident in translated texts which overcome the borders of a homogeneous cultural setting and address an international audience. For these reasons, and many others that will be proposed in the present study, research on the discourse of statistics appears to be topical in the European Statistical System. Improving the way to write and comment on statistical data will make statistical texts clear and usable by all European educated people and not by experts only.

1.2 European Statistics Code of Practice

As already mentioned, the National Statistical Institutes (NSIs) form the basis of the European Statistical System (ESS). All NSIs make their data available on their websites, paper publications and on-line publications. Data produced by NSIs are available on the Web for free. This is due to a new European policy aimed at making statistics clear and accessible to the wider public. This new trend is formally expressed by two main principles in the 2005 European Statistics Code of Practice:

> ACCESSIBILITY AND CLARITY - European statistics should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance. (Principle 15) […] Indicator 1: Statistics are presented in a form that facilitates proper interpretation and meaningful comparisons. […]

These principles have strong implications on the methods used for statistics dissemination, but also on the language of statistics. Clarity and accessibility impose upon statisticians the obligation to change the way of presenting and explaining data. This is a new requirement for specialists who usually deal with data collection and data process quality, and very little with data presentation, and now have to work in a new statistics paradigm, i.e. serving the national and the international community.

On the other hand, accessibility calls for a growing use of translation into English in order to make data accessible at international level and, especially at European level, not only for a specialized audience and the European institutions, but also for an audience of laypeople and European citizens.

An example of this approach is the new logo adopted by the Italian Statistical Institute. The traditional logo was a geometrical form – small and big (Figure 1) representing sample and unit. In 2005 the picture evolved into an open door as a visual representation of Accessibility (Figure 2).

Figure 1- Old ISTAT logo

Figure 2 – New ISTAT logo

Because of the publication of the Statistics Code of Practice, the year 2005 can be considered a line of demarcation in the development of European statistics from various perspectives. We shall analyze the way

statistics and NSIs deal with this transformation from the perspective of written language use.

For this purpose, and to lay the ground for this type of analysis, some aspects and features of the discourse of statistics will be highlighted and discussed in the following sections.

1.3 Statistics as specialized discourse

Statistics deals with very different topics (i.e. economics, social life, business, environment, and others), from the point of view of statistical analysis and interpretation. Statistics is defined by the OECD Glossary of Statistical Terms as "Numerical data relating to an aggregate of individuals; the science of collecting, analyzing and interpreting such data"[6]. Another definition relates statistics to mathematics:

> […] the science that deals with the collection, classification, analysis, and interpretation of numerical facts or data, and that, by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements[7].

Hence the discourse of statistics can be fully included in the bigger family of specialized discourse, as described by Gotti (2006), first of all because it is a "science" and therefore uses a restricted language in order to serve a circumscribed field (Firth 1957).

Some lexical, syntactic and textual features listed by Gotti (2006) as typical of specialized discourse are recognizable in statistical texts, i.e. monoreferentiality, lack of emotion, precision, lexical density, sentence length, use of the passive, anaphoric reference and textual organization.

[6] http://stats.oecd.org/glossary/search.asp (Last accessed 20 April 2012)
[7] http://dictionary.reference.com/browse/statistics (Last accessed 20 April 2012)

The texts analyzed in the present study contain these features, and they will be described and discussed by means of examples in chapter 3. The features related to specialized discourse are relevant to find out general and specific patterns of the discourse of statistics.

Studies on specialized discourse in the 20th century have provided information on specialized discourse as a specific genre also in the European environment. Within the EU context, specialized discourse is expressed in English as a Lingua Franca (ELF); its features will be discussed in the following section (1.5). Gotti (2007: 144) also refers to implications in the use of the English language within the EU, and notices that in the process of the internationalization of information:

> Not only do more and more people in Europe communicate with one another in English, but this language is the one frequently used by European Union officials to communicate with one another and with EU citizens.

This is true for statistics as well since European statisticians communicate with one another in English. Aspects relating to English communication among statisticians will be discussed in chapter 4, and the results of a specific survey on the use of the English language will be presented and commented upon in support.

1.3.1 Specialized lexis

One of the most known and evident features of specialized discourse is the specialized lexis, and obviously statistics follows this trend. Gotti (2005: 18) notices that: “There is far more than a straightforward lexical distinction at the root of specialised discourse”. However, before approaching the “far more” we prefer to start dealing with a few aspects concerning statistical specialized lexis.

Statistics is divided into three broad branches, namely business, social and demographic statistics. Within these branches, statistics uses terms from the semantic field of each specific discipline (e.g. business statistics makes a wide use of economic language). In specialized language there is a tendency to use more nouns and verbs from Latin or Greek origin for the purpose of a terminological contrast with everyday language (Gotti 2006: 40), and this approach is not alien to statistics. For example, this is the case of the verb 'increase' which is preferred to 'grow' to express the statistical trend of a certain phenomenon.

As previously mentioned, there is a specific terminology used to define data, and it implies a specific lexis which deeply characterises all statistical texts, e.g. the use of words like 'ratio', 'average', 'mean', 'vital statistics'. Specialized lexis is "not much varied to avoid ambiguities" (Gotti 2006: 26), and the lower type/token ratio of specialized texts when compared to non-specialized ones evidences this feature (see chapter 3). Type/token ratio is low also in the language of statistics (see 3.3.3) as results from corpus-assisted analysis (see chapter 3).

Wilss (1999: 81), a scholar and specialist in translation, points out that: "It is predominantly in the realm of lexis that the specialist features of such texts are located" (1999: 81). These words suggest a close connection between terminology theories and Translation studies especially when dealing with specialized discourse. This calls for further investigation to understand how the discourse of statistics differs from everyday language.

Statistical definitions, and in particular data labelling, are an interesting aspect related to specialized lexis. Labels are of crucial importance in identifying correspondence of data from different countries as well as in

describing data themselves or phenomena related to them. The label is the first reference to search for information and a sort of key word to recognise the type of data a user is searching for. Of course, other statistical issues, and not only language, are related to data labelling and definitions – namely, processes, sources, quality parameters and others. The problem of data labelling has been tackled and partially solved by the creation of international classifications which provide an exact classification of data related to a certain area, namely economic activities, education, professions, diseases. Some of the international classifications, which are in English, have an exact correspondence in other languages of the EU member states. Correspondence is clearly marked by numbers referring to each set and sub-set. This substantial classificatory work has been conducted at European level for some languages and some topics such as economic activities, which are classified by NACE (Statistical Classification of Economic Activities in the European Community). On the other hand, the International Standard Classification of Education (ISCED) was elaborated for education degrees, courses and diplomas by UNESCO, and provides a standard classification on an issue which is usually very difficult to compare at an international level, because of the different school systems. The NACE has also its Italian corresponding classification named ATECO (*Classificazione delle attività economiche* / Classification of Economic Activities). So far, it is clear that this exact correspondence can be established for some specific fields only. Classifications have proved very useful tools to facilitate statistics translation as far as terminology is concerned. Classifications are taken as international conventions which sanction the use of the same term to refer to the same data. A special

focus is on harmonization, ordering and standardization of terms, terminological editing and terminology databases.

1.3.2 Non-verbal elements in statistics

In addition to the specific lexis in statistical texts, there is a constant presence of non-verbal elements, namely tables and formula. Widdowson (1979: 52) affirms that they form the "deep structure" of specialized texts and constitute a "universal" knowledge-bank forming the basis of specialised discourse. As a matter of fact, the majority of statistical publications are composed of non-verbal elements, whereas tables and figures occupy many more pages than verbals. Statistical non-verbal elements are quite clear to the community of statisticians, even though they do not belong to the same speech community, which means that statisticians who do not share the same language but the same specialized knowledge can understand and interpret data when reported in tables or graphs. The trend to make statistics accessible to a wider public is leading to longer explanatory verbal texts because the same tables and graphs require additional explanations, namely textual explanations, when addressing lay people.

1.4 A short background on the use of English in the EU

English has not always been the main language for communication in the European Union and the European Community history. As Ostler (2010: 21) points out, "Although clearly a European language, English had, until the twentieth century, always been purely an offshore phenomenon". For centuries the British Colonies, the Commonwealth and the USA have

made English a very widespread language outside Europe. Until the Seventies, the traditional language for communication in the European institutions was French (Caliendo 2003). The European headquarters were, in fact, located in French speaking cities such as Brussels (the Commission and the Council) and Strasbourg (Parliament). In the twentieth century, French and German were indeed the most widely spoken languages and they were also taught as foreign languages for pedagogic aims. After World War II some changes started to take place: Germany and other Nordic countries shifted from French and German to English as the main foreign language taught at schools (Ostler 2010: 22); in 1973 the United Kingdom entered the European Community and since then the role of their national language has been gaining importance speeding up the already initiated trend. So far the process has kept on growing replacing the official rule of multilingualism.

1.5 Statistics and English as a Lingua Franca (ELF)

ELF (English as a Lingua Franca**)** is used for communication in Europe. In this section, aspects of ELF in the European context will be analysed and discussed with reference to the discourse of statistics. The debate on the use of ELF has developed especially in Europe during the last decade (Mauranen 2006b; Seidlhofer 2006; Jenkins 2003), therefore whoever deals with this topic handles a developing phenomenon, as remarked by Jenkins and Seidlhofer (2001)[89]:

[9] http://www.guardian.co.uk/education/2001/apr/19/languages.highereducation1 (Last accessed 7 February 2013).

> A European variety of English, sometimes labelled "Euro-English", is in the process of evolving to serve as a European lingua franca. As yet, however, this new variety of English has not been described, largely because it is at such a embryonic stage in its evolution. All we can say with any degree of certainty is that English as a lingua franca in Europe (ELFE) is likely to be some kind of European-English hybrid which, as it develops, will increasingly look to continental Europe rather than to Britain or the United States for its norms of correctness and appropriateness.

The authors here refer to ELFE as a specific variety of ELF, which characterizes communication in English within the EU. In some other references (Seidlhofer 2003) we can find EIL (English as International Language). We shall refer to ELF according to Seidlhofer's (2005: 340) view: "when English is chosen as the means of communication among people from different first language backgrounds, across linguacultural boundaries, the preferred term is 'English as a lingua franca'".

ELF is still under study, and its development grows together with the growth of the EU. However, some features of ELF are already known (see Jenkins / Seidlhofer 2001), for instance, the loss of 's' for the third person or the preference for nouns instead of verbs. ELF has been found to differ from British and US English as "a reality": "Whether pragmatic or ideological or both, this use of English as a lingua franca (ELF) *is* a reality. It declares itself independent of the norms of English as a native language (ENL)" (Seidlhofer / Breiteneder / Pitzl 2006: 4, emphasis in the original).

The need for communication among EU members has led to the use of a communication language and ELF meets the requirements needed. We can therefore speak of a "functional profile" of English in Europe (Seidlhofer / Breiteneder / Pitzl 2006: 2). The main function of ELF is to enable communication among people of different mother-tongues. The use of ELF in this light is a top-down process which from the EU

institutions reaches out to the citizens. However, we can also say that it is a bottom-up phenomenon for the generalized use of English in Europeans' everyday life, namely in music, advertisement, the Internet. European pupils study and know English whatever their education. In 2001, Eurostat found that more than 90% of pupils in secondary schools in the EU study English. English dominates also the research field, where articles and papers must be in English to address the international audience. For this reason scientists, whatever their background, and statisticians as well, can be considered members of an international community sharing a common language, which is English or rather ELF. Myers-Scotton (2002: 80) terms this the "snow-ball effect": "The more people learn a language, the more useful it becomes, and the more useful it is, the more people want to learn it".

At the level of EU institutions the situation is very similar and the use of English is now quite established and accepted, even though multilingualism is still provided by the EU as an equal right and weight to all EU member states. It can be said that at this stage the pragmatic use of language differs from EU official policy. The same happens at Eurostat level where national delegates speak in English with one another and in official meetings. Translation is provided only in Eurostat top-level meetings[10]. This transformation in the use of ELF at EU level has been already studied, and many look at it as a positive phenomenon (Seidlhofer / Breiteneder / Pitzl 2006: 3):

> Officially, the European Union, being the most influential European institution, pursues a language policy which promotes multilingualism and equal linguistic rights. Yet, when it comes to in-house communication, official policies are often discarded in order to facilitate the working process.

[10] Information obtained *via* interviews to Italian statisticians working at ISTAT.

> […] Despite widespread criticism of its dominance, it has to be acknowledged that English does serve the ideal of European integration and facilitate movement across borders.

Direct communication by means of ELF is indeed preferable to the intermediation of a translator and the use of the same language can be also a way to facilitate the construction of a stronger and shared European culture.

For all these reasons, some aspects of ELF will be discussed in the following sections, especially those concerning its use in Statistical Yearbooks.

1.5.1 ELF and multilingualism within the EU

Multilingualism within the EU means respect for national languages and is constitutive of equal status and importance for all member states. With the EU-25 and then EU-27 member states, the use of all national languages has become step by step more complex. The EU Directorate General for Translation (DGT) is one of the biggest offices of the world entirely devoted to translation activities. Still the problem of communication among different people and officers within the EU cannot be solved simply by recourse to a big amount of translation. As far as concerns spoken language people need to directly communicate with one another, and for the last twenty/thirty years English has increasingly played that role especially in Europe. As reported by scholars, only one out of four users of English is a native speaker (cf. Crystal 2003). Anyhow, the use of English among non-native speakers, as emphasized by House (2003), should not be considered a threat to multilingualism. She distinguishes between language for communication and language for identification, the former being identified with ‘Lingua

Franca'. As explained by House, the language for communication is the language used to establish a relationship of mutual exchange of information, knowledge, etc. where the main aim is communication among people who do not share the same L1; language for identification, instead, is the local language we use to identify ourselves and to present ourselves as members of a specific linguistic-cultural community. Nelson Mandela used to say: " If you talk to a man in a language he understands, that goes to his head. If you talk to him in his language, that goes to his heart."

Also in Mandela's perspective these two types of languages serve different purposes. Within the EU, where many different speech communities meet and speak, a language for communication and hence a *lingua franca* is extremely useful.

The phenomenon of a Lingua Franca has been studied from several perspectives throughout the centuries. The case of Swahili in East Africa is a clear example of this type of language, not a national language, rather a language used at the beginning of the 17th century for Arabs and East-African people to communicate with one another for trade purposes independent of national borders. In 1990 the English language was compared to Swahili by Julius Nyerere, the first president of independent Tanzania, who referred to English as the Swahili of the world (Nyerere1990). Swahili is now Tanzanian national language and has been one of the strongest means for the unity of Tanzania and its stability. The value of Swahili in constructing the unity of the people of Tanzania beyond ethnic roots is evidence of how a *lingua franca* not only serves communicative purposes but can also foster and facilitate unity in diversity. Unity of different countries does not mean deleting

national identities and differences but establishing a common ground for different people and peoples to cooperate and build new and inclusive entities, as is the case of Europe. The approach proposed by House (2003: 562) is interesting in this respect:

> […] all the official languages of the EU member states have been given equal status. With the increased number of member states, this policy is a serious problem, a problem which could be solved by adopting ELF for the EU. Once the position of English as the vehicular language were recognized, resources would be freed for supporting *all* other European languages. ELF would need to be taught intensively and early on as a true *second* language. More money and time could then be allotted for teaching and otherwise supporting other European languages (especially minority languages) in a flexible fashion, tailor-made to regionally and locally differing needs. (Italics is in the original)

In House's perspective, "ELF need not be a threat" (2003: 562), rather it can be viewed as an enrichment within the EU. And this is how we shall look at it in the present study.

The use of ELF is not just limited to spoken language as a communicative tool for people at meetings and in informal conversations. EU documents albeit official in all EU national languages are to be drafted in one language and the use of English for this purpose has increasingly spread. Murphy (2008: 21) presents diachronic data on the source languages from which documents are translated at the Directorate-General for Translation. The languages that appear to be most frequently used are English, French and German. A diachronic comparison of percentages between 1997 and 2007 can enable us to perceive the increasing use of English as the language for communication. In the following Table the percentage of translations from English, French and German are listed to clarify the trend towards the use of English as the language chosen for drafting official documents:

	English	**French**	**German**
1997	45.4%	40.4%	5.4%
2006	72%	14%	2.8%

Table 1 – A diachronic comparison of percentages of EU texts drafted in English, French and German and then translated into EU national languages in the years 1997 and 2006.

Table 1 is evidence of a trend beyond any official decision, which has led English to be the main language for communication within the EU. The reverse, i.e. translating EU documents from English into other languages, is also implemented in the drafting of the Eurostat Statistical Yearbook, which is included in the corpus used for research. The Eurostat Yearbook, which is described as "a definitive collection of statistical information on the European Union"[11] is written in English and then translated into French and German only. For the purposes of the present research its specific ELF features will be collected and discussed by means of a corpus-assisted analysis in chapter 3.

1.5.2 ELF and cultural settings within the EU

EU documents that are drafted in English use a variety of ELF, since they are written by native and non-native English speakers who are experts on a specific matter (Murphy 2008: 22). Hence, these texts are not embedded in any particular culture. The use of English is not expression of, say, British, Irish, American or Australian culture, but of EU intercultural situation as noticed by Pym (2004): "Meanings are […] in the intercultural situations in which languages are being used". ELF in this respect differs from national languages and is just partially related to

standard English (cf. House 2003). In his book, *Un Italiano per l'Europa*, Tosi (2007) points out that the function of the English language, which is used in practice though not formally mentioned by EU provisions, leads to the writing of texts in English that are not culturally marked so as to include a virtual translation into all other languages[12]. This means that the texts are indeed drafted in English, but this is a different English which is neither British nor American, it is ELF. Jenkins (2003: 1) defines ELF as "A contact language used among people who do not share a first language, and is commonly understood to mean a second (or sub-sequent) language of its speakers". She refers to spoken language, like the majority of ELF scholars (Seidlhhofer, Mauranen, House and others), but written ELF is now gaining importance within the EU, as already mentioned in this section. This aspect requires further study to recognize features and patterns of written ELF within the EU. In chapter 3, the data collected from our corpus provide some information of how ELF features are recognizable in Eurostat Statistical Yearbooks.

The EU is an intercultural setting of different nations with a European common background. European culture is not homogenously shared, we are all aware of problems related to the European Constitution, but something European is in all member-states and citizens. Our European common background cannot be ignored, our common history, our old traditions are there, Europeans belong to the same continent. Europe is

[11] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home. (Last accessed 15 December 2012)

[12] Tosi (2007: 164) writes: "L'esistenza di una lingua franca, che funziona *de facto* ma non è ammessa *de iure*, ha predisposto i servizi di traduzione ad adottare un pragmatismo quotidiano: il testo inglese poco marcato "culturalmente" deve contenere la traduzione virtuale in tutte le lingue." (italics in the original).

characterized by a multicultural setting; this expression is preferable to 'multiculturalism' because *-ism* tends to have a negative implication. As Murphy (2008: 15) remarks, "Countries adhering to the European Union are bound to become multicultural societies, which welcome peoples from different countries and backgrounds". This inclusive approach is in the EU chromosomes: "The first Community organization was created in the aftermath of the Second World War when reconstructing the economy of the European continent and ensuring a lasting peace appeared necessary."[13] Peace is the result of a cooperation policy that overcomes confrontation. And cooperation stems from the need to live and grow together.

The notion of ELF as a language for communication originates from the same need for communicating with one another and overcoming barriers. In this sense the way forward proposed by Seidlhofer (2003: 23) as "Intercultural competence achieved through a plurilingualism that integrates EIL[14] within the European context" needs to be taken into account. This approach is not against EU multilingualism; in line with House (2004), Seidhlofer remarks that both ELF and the national languages of EU member states can co-exist in favor of an enriched intercultural competence. ELF is the language that can overcome national, cultural and linguistic borders to create direct communication and to strengthen the common European identity, and, as Snell-Hornby (2000: 17) writes, is "the Europeak of our multilingual continent".

1.5.3 Written ELF

---

[13] http://europa.eu/legislation_summaries/institutional_affairs/treaties/treaties_ecsc_en.htm (Last accessed 10 February 2012)

[14] EIL – English as International Language.

As already mentioned above (Section 1.5.2), ELF has been studied mainly as spoken language. Still, the substantial growth of texts written directly in English within the EU by non-natives, as well as the huge number of publications translated into English from national languages, has drawn scholars' attention to this particular mode of written language. The first observation is that ELF is associated both to English texts written by non-natives and to texts translated into English for an international (or European) public (cf. Taviano 2010). This kind of approach which relates English translation to ELF is particularly useful for the purpose of the present study, which aims to compare three varieties of language, namely the EU English language, Irish Native English and the English language of translated statistical publications addressed to the EU area. In spite of differences some ELF features and collocations can also detected in the English translation of National Statistics addressed to the EU. The relationship between ELF and translation will be analyzed in depth later (see Chapter 3), because similarities and differences have been found to depend more on the addressee/recipient of the written text than the writer/translator.

This observation is confirmed by the corpus-assisted analysis in chapter 3 and by the data provided by the interviews to European statisticians discussed in chapter 4. Murphy (2008: 22) remarks that "texts in the European institutions have an unusually hybrid nature". This is due to the way texts are prepared, the use of previous documents and the practice of "cut and paste". In addition to that, various actors intervene at different stages in the drafting phase to change, correct, integrate, and edit texts. Even in their first drafts various hands often contribute to writing the different sections. This is the case of statistical yearbooks, where

different groups of statisticians who are experts in a specific statistical area compose the chapter they are in charge of. Moreover, Eurostat statistical yearbooks are the result of information from various European countries and regions, hence hybridization is amplified.

Once these texts have been drafted in English, they are edited by the DGT (Directorate General for Translation). Amanda Murphy (2008), who has studied the editing process compares two corpora of EU, non-edited and edited English texts. The differences she notices in the edited texts, i.e. the changes and corrections made by editors for the final edition, are quite interesting to the purpose of the present research. Just to give a few examples, an overuse of the definite article in non-edited texts (Murphy 2008: 55) can be compared to the same phenomenon in the statistical corpus, furthermore the use of ‘in order to’ and ‘as well as’, which are preferred in edited texts, could point to the feature of lexical explicitation in ELF. Lexical explicitation, which is recognized as an ELF feature, refers to an effort by editors and translators to make explicit what could seem to be implicit and not clearly understandable. Explicitation is also included among ‘translation universals’ (Baker 2001), therefore this and other features which are related to ‘translation universals’ will be discussed in section 1.8 of the present chapter.

Some specific collocations characterize ELF, for example the use of adjectives and past participles in post-modifier position (Taviano 2010: 29); or the large use of nominalization which prefers nouns to replace verbs (Taviano 2010; Murphy 2008); some features of American English (e.g. ‘z’ spelling and others), which are (somewhat surprisingly in Europe) preferred to the British ones. These and other patterns which have been identified as typical of ELF can also be found in the statistical

language of Eurostat and of translated texts. Some similarities between Eurostat and translated statistical texts are evidenced by the data analyzed. In some aspects we could speak of ELF both for statistical texts translated into English and addressed to an international audience and for Eurostat English. We could assume that the use of ELF is also related to the recipient, still this interpretation needs further analysis supported by data (see chapter 3). This would mean that when an English text is intended for an international public then some ELF features and strategies are also adopted by English native speakers who are in charge of the translation.

1.6 Statistics and communication

Communicating by means of ELF, which enables direct exchanges, is not free of obstacles if House (1999) titles her study "Misunderstanding in intercultural communication: interactions in English as a lingua franca and the myth of mutual intelligibility". The main hypothesis of her study is that misunderstandings in ELF talks are due to different frames of the cultural knowledge based in L1 and to interactional norms. In this way House (1999) supports the idea that an effective communication is also influenced by cultural backgrounds, and not by language proficiency only. What she observes referring to talks can also be referred to the written language of European national statistics when national publications are addressed to an international public by means of translations into English. In this way publications on specific national data are proposed to the international arena, specifically the EU, though with no specific attention to the readers' different cultural backgrounds.

If, for example, we look at the case of migration, we cannot interpret data in the same way for all countries. The following examples (1), (2) and (3) show three different situations where further cultural knowledge is necessary in order to interpret data appropriately:

(1) Between 1992 and 1994 only the migration of citizens of the Republic of Slovenia is included. According to legal regulations of the Republic of Slovenia concerning personal conditions, the ex-citizens of the SFR Yugoslavia who did not accept or fulfill conditions for acquiring citizenship of the Republic of Slovenia became foreigners. (*Slovenia 2007*)[15]

(2) In 2007, 8,442 people were granted Hungarian citizenship, while this number was 6,172 in the preceding year. 72% of citizenship recipients were former Romanians and 9-10% Serbians and Montenegrins and Ukrainians. (*Hungary 2008*)

(3) Immigration is often thought of as immigration of foreign citizens, but Danes can also immigrate. 31 per cent of all immigrants are Danish citizens returning after a shorter or longer period abroad or who are born by Danish parents abroad. (*Denmark 2009*).

In (1) the reader should be informed of the terrible war which affected former Yougoslavia as well as the new and difficult political balance established after 1992 by the Dayton peace agreement among different countries and peoples. Such details are necessary in order to fully understand the reported information. In (2) the migration of a high number of Romanians is due to the contended Transilvania land claimed by Hungarian people. This is a key detail to understand the difference between Romanian migration to Italy (for example) and to Hungary, the former due to economic problems the latter to historical and nationalist reasons. In (3) the explanation provided on the way of counting migrants would help to understand a specific Danish lifestyle.

[15] In these three examples only, the country and year of the source National Statistical Yearbook are reported in order to emphasize cultural differences.

Communication gets even more difficult when the subject to disseminate is statistics and therefore what is needed is specialized communication. A definition of specialized communication is provided by Shubert (2011):

> Specialized communication comprises purposeful, informative, monolingual and multilingual, oral and written communicative acts of a specialized content carried out with optimized means of communication by agents pursuing their professional duties.

In these few lines Shubert gives a picture of the complexity of such an activity as specialized communication. Studies on Language for Special Purposes (LSP) have developed in line with specialized communication which Shubert (2011) also interprets as a joint action between LSP studies and translation, referring both to monolingual and multilingual communicative acts. The dissemination of statistics throughout the EU falls in this field.

Statistics, like any other specialized discipline, has its own language and definitions which are not easily understandable by non-specialized readers. In the last decade an effort to disseminate statistical information to a larger public has impelled reflection and actions on the part of statisticians as part of their "professional duties"(Shubert 2011). This issue is related to dissemination both by means of printed publications and on the Internet. Since 2005, when the Statistics Code of Practice was passed, some proposals have circulated on how to achieve better communication. Attention has been given to disseminating statistics also *via* the media, and not only by means of specialized publications or conferences. The question was how to make statistics meaningful and clear to the audience (i.e. readers and listeners). For this purpose the

United Nations Economic Commission for Europe (UNECE) has published a guide titled *Making Data Meaningful*[16]. This guide which is divided into three parts, is a very useful attempt to help statisticians in the difficult task of presenting data. The first volume, published in 2006, provides guidelines and examples on the use of effective writing techniques to make data meaningful. The second volume (2008) provides guidelines and examples to prepare effective tables, charts and maps, and on how to use other forms of visualization to make data meaningful. The third volume (2011) mainly focuses on communicating with the media. A specific publication was also prepared by the OECD (Organization for Economic Co-operation and Development), i.e. the *Innovative Approaches to Turning Statistics into Knowledge* (2008)[17]. *Statistics Denmark* and *Statistics New Zealand* were forerunners in this field, presenting a study on the dissemination of statistics based on feedback by users, *Good Dissemination Practice in Statistics New Zealand and and Statistics Denmark* (2003)[18]. These texts provide guidelines and tips on best practices for disseminating statistics including suggestions on the writing style to catch the reader's attention and interest. A useful example is provided by *Making Data Meaningful* (2006):

> For data to be meaningful to a general audience, it is important to find meaning in the numbers. The word "story" often alarms people in the statistical/scientific world, because it has overtones of fiction or embellishment that might lead to misinterpretation of the data. This view might be justified if analysts do not approach the data with care and respect. However, the alternative, i.e. avoiding a story, may be far worse. People often distrust statistics and feel they are misleading, because they cannot understand the data. This occurs because we, the people who produce data,

[16] http://www.unece.org/fileadmin/DAM/stats/documents/writing (Last accessed 1 June 2012)
[17] www.oecd.org/oecdworldforum/statknowledge (Last accessed 1 June 2012)
[18] http://www.dst.dk/pukora/epub/upload/6841/gooddissem.pdf (Last accessed 1June 2012)

> fail to make them relevant and explain them in terms that people can understand. Without a story line, a release becomes just a simple description of numbers.
> Do not burden the reader with too many numbers in the body of the text and use only key rounded figures. Less important numbers should be relegated to accompanying tables. Use the text to present analysis, trends and context, not to repeat values in the tables.

This extract suggests that statistics should be looked at from the reader's point of view. In the guide, emphasis is on the relevance of readability and intelligibility rather than extreme correctness: "use only key rounded figures". Another aspect which is relevant to the aim of the present work is the explicit tip to use "text to present analysis […] not to repeat values in tables". The verbal statistical "text" is considered, and is indeed, understandable by all educated readers with no need for specific statistical background. A text can well express the relevant changes or differences among numbers which could not be easily detected by non expert readers.

The general commitment of European statistics towards increased accessibility and clarity is also testified by a number of courses on statistical communication offered by Eurostat to statisticians from EU member states[19].

1.7 Statistics and language

This section offers some observations on the relationship between language and statistics.

Statistical definitions rely on language. An easy-to-make observation is that even-though Eurostat defines National Institutes as 'National

[19] http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/ESTP_Programme2012.pdf (Last accessed 5 February 2012)

Statistical Institute'(NSI), different names are still found on official websites. An example of this is the EU immigration portal,[20] where Italy refers to its National Statistical Institute as 'Italian National Statistical Institute', Spain as 'National Institute of Statistics' and Portugal 'Statistics National Institute'. This means that the same national statistical institution, with similar tasks and aims, is named differently by European countries when using the English language. This fact could induce non expert readers, such as migrants, who access the webpage to believe that they are different kinds of institutions, which is not true.

As already mentioned, the language of statistics has not been much investigated so far. Its cultural implications lead us to reflect on the problem raised by Gotti (2006), even though in a completely different context referring to legal discourse, and which could be applied to the statistical domain: when terminology is very culture-bound a satisfactory translation of all terms in one text is at times impossible. Statistical definitions and their translation are still a field to be investigated more deeply not only as regards interrelations among different languages but also as different cultural products. Statistics are used to describe societies, lifestyles, and many other aspects related to phenomena which are deeply interconnected with the cultural setting from which they are produced.

Another branch of the discourse of statistics which still needs to be further investigated deals with metadata and comments. Metadata refer to the description of data source, statistical processes and statistical quality requirements. Comments on the other hand provide a description of data comparing them in time (example 4) and space (example 5):

[20] http://ec.europa.eu/immigration (Last accessed 20 February 2012)

(4) On the basis of the most recent projections on population development, the demographic dependency ratio will increase to 0.92 in 2010 and reach 1.1 in 2030. (*Transtat*)

(5) Around one third of employees in the EU-27 participated in continuing vocational training (CVT) courses during 2005. Among the Member States, the proportion ranged from 50 % or more in the Czech Republic and Slovenia to 15 % or less in Greece, Lithuania, Latvia and Bulgaria. (*Eustat*)

Example (4) shows a typical way of commenting data in time (diachronic comparison) also using projections to compare present and future, whereas example (5) shows a statistical comparison in space. It is typical of Eurostat to compare data from different EU regions, namely, Czech Republic, Slovenia, Greece, Lithuania, Latvia and Bulgaria. Comments and metadata are included in the textual part of statistical publications and releases. Glossaries are sometimes included as well, in order to provide explanations on specific terms (example 6), or on general terms used in a specific way (example 7):

(6) **Infant mortality**: deaths of live births between birth and exact age one year; deaths before registration are included.

(7) **Illiterate:** without primary school certificate (including people holding a special certificate having attended the third grade of primary school); people who can read or write; unable to read or write.[21]

Glossaries are very helpful to understand data reported in statistical publications and to clarify what and whom they refer to. As shown in example (7) the definition 'illiterate' could be interpreted in different ways, for instance, only referred to people unable to read or write; in this case, instead, it refers to a broader group of people whose main characteristic is that they have no 'primary school certificate'.

[21] http://www.istat.it/it/archivio/17969#G (Last accessed 20April 2012)

Therefore, glossaries can increase clarity standards. This is also confirmed by the results of the survey discussed in chapter 4.

1.8 Language and quality in statistics

Before analyzing the relationships between language and quality in statistics we should focus on the meaning of 'quality' in the statistical domain. To this purpose, we shall report some definitions from different sources. The first is the definition of 'National quality' by OECD:

> Data quality relates to information about sampling and non-sampling errors, as well as associated statistical reporting and adjustments intended to quantify and account for these errors. There are both direct and indirect measures of data quality. Direct measures deal with the survey itself, while indirect measures are the result of process evaluations or comparative studies.[22]

We can clearly understand that data quality is very much related to the data process and to the reduction of errors, therefore data quality is intended as correctness of and control on the various stages for processing data in order to collect trustworthy statistical information.
ISTAT has elaborated a specific system to grant data quality; it is presented in the official website as follows:

> **Data quality**
>
> SIQual, the information system on quality, contains information on the execution of Istat primary surveys and secondary studies and on activities developed to guarantee quality of the statistical information. The system describes the production process and its characteristics: information content; phases and operations of the production process; activities to prevent, monitor and evaluate errors.[23]

[22] http://stats.oecd.org/glossary/detail.asp?ID=2217 (Last Accessed 10 December 2012)

The information contained in the OECD corresponds to the information provided by ISTAT, both identify error prevention and process control as the main aspects of quality in statistics. However, when we move to Eurostat's definition of 'quality' we can notice a different perspective which is quite interesting to the purpose of the present study. Eurostat defines quality in statistics by means of six criteria[24]:

- relevance;
- accuracy;
- timeliness and punctuality;
- accessibility and clarity;
- comparability and
- coherence.

Eurostat's criteria on 'quality in statistics' do not only refer to accuracy, coherence and comparability, which are all aspects of 'quality' directly connected to data and their truthfulness, or to the methods implemented by statisticians to process them; Eurostat's criteria also include 'timeliness and punctuality', and 'accessibility and clarity'. These two criteria are deeply innovative and rely on the user's perspective to assess data quality. So far quality has been always related to methodology, as also confirmed by the quoted definitions (OECD, ISTAT). By including these new criteria Eurostat affirms that assessing 'quality' is not circumscribed to the scientific community of statisticians. Tests on 'clarity and accessibility' especially rely on users' assessment, therefore 'quality in statistics' is no longer a matter to be discussed exclusively

[23] http://www.istat.it/en/tools/data-quality (Last accessed 10 January 2013)
[24] http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/ess%20quality%20definition.pdf (Last accessed 3 December 2012)

within the statisticians' scientific community, but both expert and non-expert users are involved.

In 2008, a scientific study on the use of the English language in statistical discourse was accepted for the first time at the International Conference "Q2008" on Quality in Statistics held in Rome.[25] The Conference, which is held every two years, gathers top-level statisticians from all over the world to discuss on 'quality in statistics'. The large majority of participants are statistical methodologists. For this reason the inclusion of a paper on the use of the English language in the dissemination of statistical knowledge has been an important recognition. The paper presented some proposals for a better English translation of editorial products, and for achieving clarity and accessibility to meet users' needs.

This link between language and quality in statistics was again emphasized in 2009 by Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 (Art. 12. Statistical Quality) which states that:

> 1. To guarantee the quality of results, European statistics shall be developed, produced and disseminated on the basis of uniform standards and of harmonised methods. In this respect, the following quality criteria shall apply: (e) 'accessibility' and 'clarity', which refer to the conditions and modalities by which users can obtain, use and interpret data; [...]

The introduction of this new Regulation acknowledges addressees' perspective in statistics dissemination, and clarity and accessibility as quality standards. This Regulation is the result of a long legislative route, because clarity in writing is an emerging concern in EU institutions in general and not only at Eurostat level. The "Fight the Fog Campaign" has

worked on this field from the end of the 90s and has drawn the institutions' attention to this specific topic which has often been left aside. In 2010, the EC Directorate General for Translation published a booklet on *Clear Writing*, which, in the same line as the "Fight the Fog Campaign", offers and illustrates ten tips for achieving clarity. The ten tips are reported below:

> **Tip 1: Think before you write** Clear writing starts with clear thinking. Ask yourself: Who will be reading the document? What are you trying to achieve? What points must the document cover?
> **Tip 2: Focus on the reader** — be direct and interesting Try to see things from the point of view of your readers. Involve them. Imagine which questions they might ask. Interest them.
> **Tip 3: Get your document into shape** Give your document the right structure and avoid mistakes commonly made at the Commission.
> **Tip 4: KISS: K**eep **I**t **S**hort and **S**imple Don't be afraid to go for the shorter option. Avoid over-long sentences.
> **Tip 5: Make sense** — structure your sentences Arrange ideas in logical (often chronological) order. Don't bury important information in the middle of the sentence.
> **Tip 6: Cut out excess nouns** — verb forms are livelier. Avoid noun disease by using verbs and verbal forms instead.
> **Tip 7: Be concrete, not abstract -** Concrete messages are clear — abstract language can be vague and off-putting.
> **Tip 8: Prefer active verbs to passive** — and name the agent. If you change passive verb forms to active ones, your writing will become clearer because you will be forced to say who is responsible for the action.
> **Tip 9: Beware of false friends, jargon and abbreviations -** We all know how and why it happens, but many say this is the cardinal sin of Eurocratic writing.
> **Tip 10: Revise and check -** Don't just rely on your spelling chequer!
> (Note on the explanation of Tip 10: Yes, 'chequer' is a deliberate mistake.)[26]

All these tips are very useful to the drafters of statistical publications, and to translators as well. For example, choosing verbs instead of nouns, or

---

[25]Patrizia Collesi - *Proposing Editorial Catalogue and Presenting Editorial Products in English: Some Proposals for Better Translations having in Mind Users' Needs.* http://www.istat.it/istat/eventi/q2008/sessions/35.html (Last accessed 3 November 2011)

[26]http://ec.europa.eu/dgs/translation/publications/magazines/languagestranslation/documents/issue_01_en.pdf (Last accessed April 2012)

active verbs in the place of passive ones, are useful suggestions also in the translation process, when the aim is to be clearer. Translators want to be clear just as "Editors strive for clarity, and use any linguistic means they can to make the text clearer" (Murphy 2008: 83). Clarity in presenting and disseminating statistics is a challenge for statisticians and editors; this concern should also involve translators who are mediators between the statistics national product and the international users. 'Clarity and Accessibility' as parameters of quality in statistics should be preserved also in English translation. That is why in the following section we shall discuss issues related to the translation of statistical documents.

1.9 Statistics and Translation

English was chosen *de facto* (Tosi 2007) as the language for communicating and disseminating statistics at the European level. To this purpose, English translations of statistics national texts have been fostered by the EU. At the end of the Peer Review at ISTAT in 2006, among other comments, we can read: "It is also necessary to assist users in understanding statistics on the Web. Development of an English version of the documentation is important as this is more or less non-existent nowadays."[27] This comment is in line with the "Accessibility" principle and opening the statistics world to the international public.

NSIs are required by the EU to produce an English version of their most important output and publications. In order to meet this requirement the majority of member-state NSIs present their data in the national language

[27] http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/PEER_REVIEW_IT_2006/EN/PEER_REVIEW_IT_2006-EN.PDF (p.11) (Last accessed 3 November 2010)

and provide an English translation of their most relevant publications. Translation into English is one of the means adopted by NSIs to make their data accessible abroad; therefore the effort of translating their products should be included in the framework of communicating national statistics to an international audience. Communicating with an international public– specifically, European national institutions and citizens – deals with cultural differences which cannot be "overlooked" as observed by Kastberg (2007: 1) who points out that "cultural issues are inherent in technical texts and should not be overlooked, both in translation practice an in translation training."[28]. Translators of EU-member-states statistics handle a very difficult task. Their work is to translate into English from whatever national language in order to produce texts that can be readable both by native-English speaker and by – this is the largest group – L2 English speakers. Sandrini (2006), when discussing translation in relation to globalization and different cultures, writes:

> Translation is text production for another – relative to the source text – linguistic background. Translation studies have stressed the fact that language is an integral part of a national culture and that consequently there is no language transfer without the impact of cultural factors. Translation thus is the dissemination of specialized knowledge in another linguistic and cultural context.

Cultural factors are always to be taken into account when translating a text, even more when this text is not addressed to a specific national community with one language and one culture. The question is how to disseminate a specialized knowledge characterized by national and culture-bound features to people who belong to different cultural settings

[28] http://www.jostrans.org/issue08/art_kastberg.pdf (Last accessed 3 November 2012)

by means of a language (English) which represents neither the source culture neither the target one. We do not have an answer to this question, however the present study aims to offer comments and suggestions to deal with the translation of statistical texts in a renewed perspective.

"Clarity and accessibility" criteria are to be met also by translated statistical publications. To this aim we can refer to 'translation studies' which have investigated some shared features among all translated texts. It has been noticed (Baker 2001) that translated texts are in general more explicit than original ones. This effect is due to translators' mediation in their effort to be understood, and to render explicit what they consider implicit in the text. This specific feature has been called 'explicitation' and included among the features of "translation universals" (Mauranen / Kujamäki 2004). The notion of some universal features shared by all translated texts is an evolution of Blum-Kulka's hypothesis (1986). It means that no matter the source and target languages there are universal features and strategies that are used by translators when drafting translations. By using corpus linguistics and drawing from translation studies, Mona Baker (2006: 176) describes the distinctive features of translations supporting the theory of Translation Universals, namely simplification, explicitation, normalization or conservatism and levelling out. She describes them as follows:

> simplification –the idea that translators subconsciously simplify the language or message or both, explicitation – the tendency to spell things out in translation, including, in its simplest form, the practice of adding back ground information and normalization or conservatism – the tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them (ibid. 176)

Simplification and explicitation are very interesting in the perspective of 'clarity and accessibility', because they make the translated texts more explicit, simpler and we could infer clearer. The approach proposed by Baker has found some opposition. Chesterman (2004: 43) analyses it as part of the descriptive route of translation studies. He argues that

> […] any claim for translation universals can really only be an approximation. But this doesn't matter, as long as scholars are aware of what they are claiming. After all, what these corpus scholars are basically doing is seeking generalisations. We seek generalisations that are as extensive as possible. Less-than-universal claims can still be interesting and valuable. Any level of generalisation can increase understanding.

Even though Chesterman disagrees with the theory of "Translation Universals" he gives us a reason for examining it in detail when he writes that "generalisation can increase understanding". In the present study we shall analyse some aspects of generalisation which are meant to "increase" understanding with reference to the discourse of statistics in the European context. Also some of the 'Translation universals' have been detected in the corpus.

Since the first studies by Mona Baker (1996) translation universals have been discussed taking into account the socio-cultural constraints and other contextual variables in which translations and source texts are produced. The corpus-based approach is indeed very useful also to analyse the translated language of statistical texts. Silvia Bernardini and Federico Zanettin (2004: 51) suggest in this respect that there is a need "to set up corpus resources so as to allow multiple comparisons across sub-corpora, such that each component can be used as a control for the mirror one".

This kind of approach has been adopted in the present research. Indeed, multiple comparisons across sub-corpora enable us to identify specific features related to translation universals. The more specific the sub-corpus is, the more its features can be studied in depth. Through this method culture-bound features can be identified and considered separately from translation features, and specific features of Eurostat language can be identified as different from native and translated statistical texts.

The concern for statistical translation is relatively new, and is linked to the grown interest in statistical information. In order to translate statistics, NSIs need to rely on translators who have "the knowledge, the competence and the recognised status of an expert" (Snell-Hornby 1992: 10).

Here we enter the field of 'domain expertise' as proposed by Adab (2000) who states that effective communication through translation into the second language is achieved, given that users of ELF "may share domain-specific expertise". Jan Engberg[29], professor at Aarhus University and expert in specialized communication, also affirms that ELF is not enough to communicate specialized knowledge when the communication takes place among people who do not belong to the same discourse community i.e. when they do not share the same domain expertise. The challenge is for statisticians, and to some extent for translators of statistics, to communicate with a wider public and to make texts readable and understandable to non-expert users, aiming at clarity and accessibility as required by the Statistics Code of Practice.

[29] Seminar at "Federico II" University - *Analysing Knowledge Elements in Specialized Texts*- 16 October 2012.

Statistical translation into English is addressed to an international public and not to English native speakers as it is for other specialized translation (e.g. home appliances instruction leaflets, which provide translation in several languages). The use of English as a means of communication at international (in our case European) level makes the language more similar to ELF than to native language standards.

In her book *Introducing Corpora in Translation Studies* (2004) Maeve Olohan provides a long and articulated review of Translation Studies research, reporting evidence of a new trend, i.e. a "scientific" approach to the study of translation. It is no more the deterministic approach in which "meanings are objective and stable […] the translator's job is to find and transfer these [meanings] and hence to remain as invisible as possible" Chesterman and Arrojo (2000)[30].

Translators, on the contrary, influence very much readability and clarity in a text, and Mona Baker (2006: 1) highlights the role played by translation at an international level:

> In this conflict-ridden and globalized world, translation is central to the ability of all parties to legitimize their version of events, especially in view of the fact that political and other types of conflict today are played out in the international arena and can no longer be resolved by appealing to local constituencies alone.

Internationalization has increased the need for translation. However, Baker also reminds us of the delicate role played by translation and translators. In this sense also data and statistics play a political role, as described in the introduction of the present research, and EU-member

[30] Chesterman and Arrojo (2000: 151) oppose this approach to non-essentialism where "meanings […] are inherently non-stable, […] they have to be interpreted in each individual instance, and hence […] the translator is inevitably visible".

states offer a sort of national identity card of their credibility when they deliver statistics to other countries.

In the last two decades translation studies have moved on to investigate the translated text as such:

> In Gideon Toury's conceptual map of Translation Studies, the transition from theoretical and descriptive study to translator training, translation aid and translation criticism is not direct but occurs through the establishment of "bridging rules" by practitioners. Corpus-based research into the universals of translation is strengthening the pivotal role of description in translation studies through the development of an explicit, coherent methodology and the acquisition of new knowledge about translational behavior, without necessarily paying attention to such bridging rules" (Laviosa 2008: 119).

The present study refers to translation studies focusing on 'Translation Universals' and translation as international communication. Translation addressed to people who do not belong to the same speech-community involves specific communication strategies (cf. Nunn 2005), and this is also what this study aims to investigate.

1.9 Final remarks

In this chapter we have discussed some features of the discourse of statistics which have been adopted in the present research to describe and to analyze this type of discourse in the English language. Features of specialized discourse and ELF as well as aspects of translation have been detected and will be analysed in chapter 3 by means of data. Here we can observe that ELF within the EU context is used both for drafting and translating statistics texts. It can be noticed that some features, for instance 'nominalization', are shared both by EU texts and the English translations of statistical texts of European member states. This is a counter-tendency to

clarity (Tip 6), but it is one of ELF features occurring in both types of writings. Translation features which have been discussed in the last section confirm that clarity, explicitation, translation and ELF are interconnected in the EU effort to disseminate statistics to the European citizens.

2. METHODOLOGICAL FRAMEWORK

This chapter presents the methodological approach used for the present research. It contains a description of the aims of the research and provides an overview of the EU policy adopted for statistical publications in order to explain why National Statistical Yearbooks have been included in the corpus. It also describes the preliminary steps for building the corpus and the three subcorpora it is composed of as well as the methods used to analyze the corpus and compare the three subcorpora.

The last part is devoted to introducing the survey carried out for testing the statisticians' perception of language-related issues in statistical publications.

2.1 Aims

The research study investigates some aspects of the new statistical user-friendly trend, which includes expert and non-expert users feed-back in statistical quality assessment. The research, far from giving exhaustive responses on the matter, offers a circumscribed analysis of a corpus of statistical texts.

It is only in the last decade that the debate on statistical language has found room in the scientific statistical community. Thus, the aim of this research is to provide data in order to analyze specific features and patterns occurring in English statistical discourse, and to assess clarity and accessibility especially with respect to an international audience. Special attention is given to the English translations of statistical

publications which are the main vehicle, currently implemented by NSIs, to reach out the international community. This aspect falls in the field of specialized translation; as regards statistics, the field has been so far under-explored.

The research contains a corpus-based analysis of statistical texts (chapter 3); this is followed by an overview of the relationship European statisticians have with their textual products and some possible suggestions on how to improve clarity and accessibility in statistical texts (chapter 4).

The research is conceived of as a contribution towards improving statisticians' awareness of the language they use in data dissemination. It raises questions on how to address the international public and offers a picture of the state of the art. It is an attempt to describe features and patterns of the discourse of statistics. Since the user's needs is the perspective from which we read statistical texts, the focus will be on the international community as the recipient of national and European statistical publications in English.

The main research questions underlying the research are the following:

1) What are the general features of the English language of statistics used in the texts produced by Eurostat and EU member countries?
2) What are the different features and patterns that characterize texts translated into English from the texts drafted by English native speakers in their own language?
3) Do ELF features characterize *Eustat*?
4) Can ELF features be detected in *Transtat* ?
5) What are the main differences between *Eustat* and *Natstat*?

6) Do Statistical Yearbooks in English meet the requirements of Accessibility and Clarity?
7) How do European statisticians relate to data dissemination in English, and what is their perception of Accessibility and Clarity?

2.2 European statistical publications

A preliminary survey on European Statistical publications in English was carried out in order to select the most suitable ones for inclusion in the corpus.

Each European NSI has a certain number of publications which refer to different topics and also to different time-spans. The majority of publications are on specific topics such as labor market, prices, household consumption, agriculture, etc. Some other publications include data at regional and department level and refer to a specific year or to a number of years reporting the phenomenon development (time series). Others are quarterly reviews and mainly deal with economic issues. The results of agriculture, population or other censuses are published in large volumes and typically refer to the last decade. Then there are general publications which deal with various topics and refer to specific-time spans. Some publications are made of tables only, some others are accompanied by methodological information, glossaries and comments.

For the purpose of this study, only the NSIs of 25 EU member states were taken into account.

The years 2005 and 2010 mark the beginning and the end of the corpus. The starting point was marked by the adoption of the Statistics Code of Practice, which was a mile-stone in European statistics user-oriented

policy and gave a strong input to the production of statistics in English. The end was marked by the publication of data on 2010 by NSIs at the beginning of 2011. Since Romania and Bulgaria joined the European Union in 2007 they were excluded from the collection.

An .xls table with the 25 member states was prepared taking into account the availability of English publications for each country. The first data to be collected was whether all countries presented an English version of their website, and in fact they did[31]. Then the English publications common to most websites were analyzed. Three kinds of publications which were common to all national official websites were taken into account, namely the National Yearbook, *Country in Figures* and press releases.

Country in Figures is a small booklet with a short introduction and tables on the main national data. It is published in English by Eurostat and 12 EU countries. Other booklets may have slightly different titles (*Minifacts*

[31] http://www.mof.gov.cy/mof/cystat/statistics.nsf/index_en/index_en?OpenDocument (Last accessed September 2010)
http://www.dst.dk/en, www.statistik.at/web_en/ (Last accessed September 2010)
http://statbel.fgov.be/en/statistics/figures/, http://www.stat.ee/en (Last accessed September 2010)
http://www.stat.fi/index_en.html, http://www.insee.fr/en/default.asp (Last accessed September 2010)
https://www.destatis.de/EN/Homepage.html;jsessionid=1B00A9F668E59ED43BA80D27F9A948AC.cae2, http://www.statistics.gr/portal/page/portal/ESYE (Last accessed September 2010)
http://www.cso.ie/en/, http://www.istat.it/en/, http://www.stat.gov.lt/en/ (Last accessed September 2010)
http://www.csb.gov.lv/en (Last accessed September 2010)
http://www.statistiques.public.lu/en/actors/statec/index.html (Last accessed September 2010)
http://www.nso.gov.mt/, http://www.cbs.nl/en-GB/menu/home/default.htm, (Last accessed September 2010)
http://www.stat.gov.pl/gus/index_ENG_HTML.htm (Last accessed September 2010)
http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_main&xlang=en (Last accessed September 2010)
http://www.statistics.gov.uk/, http://www.czso.cz/eng/redakce.nsf/i/home (Last accessed September 2010)
http://portal.statistics.sk/showdoc.do?docid=359, http://www.stat.si/eng/index.asp (Last accessed September 2010)
http://www.ine.es/en/welcome_en.htm, http://www.scb.se/default____2154.aspx (Last accessed September 2010)
http://www.ksh.hu/?lang=en, (Last accessed September 2010)
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ (Last accessed September 2010)

about Estonia, Poketbook Germany, Women and Men in Ireland) but contain the same kind of information and follow the same structural pattern (i.e. tables without comments).

Press releases are presented in English by seventeen countries out of twenty five; the text is usually very short and in some cases, such as Germany or Estonia, only tables are given in English.

National Statistical Yearbooks are, among general publications, those which contain longer texts. Their English versions are available online for fifteen countries and Eurostat. These reasons concurred to the choice of national statistical yearbooks as texts to investigate.

As for copyright, any information available on EU member state statistical websites can be freely used for research purposes provided the source is quoted. As far as concerns data from EU member state statistical websites in the present study, no obligation to quote the source is imposed since only texts written to introduce or explain tables containing such data are reported.

2.3 National Statistical Yearbooks

A National Statistical Yearbook is a "traditional" comprehensive collection of data[32] which is published annually by NSIs and contains data referred to a specific year and divided into chapters according to the topic. It is a large volume that ranges from 300 to 700 pages. It includes many tables, all chapters are introduced by explanations and comments accompanied by methodological notes and glossaries for specific terms.

---

[32] The first publications in Europe date back to the end of the 19th and beginning of the 20th century.

All National Statistical Yearbooks have the same structure and aim, which is to provide a huge amount of data on a specific country at a specific time. National Statistical Yearbooks are a reference point and describe the overall profile of European countries from a statistical point of view. Usually, every year each National Statistical Institute proposes the same pattern of publication with up-dated information and comparison to previous data, together with projections for the future.

Eurostat also publishes a yearbook which has a similar pattern to national ones, and collects data from EU member countries under the same volume. These data are divided on the grounds of topics which allow comparisons between countries and regions. The Eurostat yearbook is written by English native speakers and is available on the Eurostat website[33] also in German and French, the latter being translations from English. All the texts included in Eurostat yearbooks are not translated from the original national language but are written directly in English.

Statistical Yearbooks can be compared since all of them are homogeneous in their structure, content, aim and users (Baker 2001: 60). They all include an introduction providing general information and contacts for the National Statistical Institutes. Each chapter focuses on such statistical topics as environment, agriculture, vital statistics and so on, and begins with an introduction on the topic, methodology and some general comments on data presented in the tables. Yearbooks are not intended to deepen the knowledge of any specific topic or interpret any aspect of social and economic life but to provide up-dated information and comparisons in time and space. They are, therefore, a sort of picture

[33] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ (Last accessed 4 September 2010)

of the country with a huge number of data and are addressed not only to statisticians but to all scholars, students and users in general who need statistical information for different purposes.

2.4 Corpus design and methodological approach

The aim of this research is to design an English comparable corpus ECC following Baker (1995: 234), hence including both translated and non-translated comparable texts. Sinclair's (1991: 20) remarks on corpus creation have guided the work from the first steps:

> The beginning of any corpus study is the creation of the corpus itself. The decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently. The results are only as good as the corpus.

The methodology adopted for the corpus design was provided by Bowker and Pearson's (2002) *Working with Specialized Language - A practical guide to using corpora*.

The entire .pdf file of each yearbook was downloaded. In some cases, such as for the Czech Republic, each chapter had to be downloaded separately since each one was a different pdf. In the case of Estonia, Italy, Lithuania, Portugal and Slovenia the yearbooks are published in English with facing-page translation. All yearbooks have tables, figures and footnotes.

After completing the downloading, all files were saved in .txt format.

In the new files all tables, figures and footnotes were deleted. Footnotes typically indicate the source of data and therefore were not interesting for the purpose of the study. Also the Contents pages were deleted as well as

all parts of the text written in the national language in the case of facing-page translation. Each file was named with its national domain and the Yearbook year (i.e. IT2009, PT2008, ND2007, etc.). The whole corpus was named ENSY (European National Statistical Yearbooks).

2.5 ENSY composition

As already mentioned the time covered by the research is 2005-2010 . For each country the last three yearbooks published in English and available online were included. The corpus consists of the English translation of European National Statistical Yearbooks from:

COUNTRY	**YEAR/S OF REFERENCE**
Cyprus	2005-2006-2007
Czech Republic	2009
Denmark	2007-2008-2009
Estonia	2008-2009-2010
Finland	2005-2006-2007
Hungary	2006-2007-2008
Italy	2007-2008-2009
Lithuania	2010
Poland	2007-2008-2009
Portugal	2006-2007-2008
Slovakia	2005-2006-2007
Slovenia	2007-2009
The Netherlands	2006-2007-2008

Table 2 – National Statistical Yearbooks translated into English included in ENSY.

All the above-mentioned files were grouped in a subcorpus named *Transtat* (Translated statistical texts). Then, following the same procedures, four yearbooks from Ireland (2007, 2008, 2009, 2010) were collected and grouped in the *Natstat* (Native English statistical texts) subcorpus. UK yearbooks were excluded because the latest UK National Statistical Yearbook was published in 2005 and has a very different structure providing a very long description of the country life and policy but few data. The new yearbooks published by UK Statistics are at regional level and do not cover all the topics included in other national Statistical yearbooks.

Another sub-corpus was created for Eurostat yearbooks (2006/7, 2008, 2009, 2010) and was named *Eustat* (EU statistical texts).

The ENSY corpus and the three subcorpora are composed as is described in the following Table:

Transtat	34 files	730,363 tokens	15,280 types
Natstat	4 files	80,259 tokens	3,677 types
Eustat	4 files	332,326 tokens	9,837 types
ENSY	**42 files**	**1,142,948 tokens**	**18,806 types**

Table 3- ENSY Corpus composition.

Eustat was separated from *Natstat* and *Transtat*, even-though written directly in English and not translated from any other language. However, it was not considered native because it is not embedded in a particular culture, rather in a multilingual setting (see section 1.5).

Eurostat production was considered more in the field of ELF (see chapter 1), according to Firth's (1996: 240) interpretation:

> Although this does not preclude the participation of English native speakers in ELF interaction, what is distinctive about ELF is that, in most cases, it is a 'contact language' between persons who share neither a common native tongue nor a common (national) culture, and for whom English is the chosen foreign language of communication.

This can be said of Eurostat as the statistical body of the EU, where English native speakers take part in interactions among non-natives who have chosen English as a foreign language for communication.

Another reason why *Eustat* was separated from *Natstat* is that the information used by Eurostat to draft the yearbook is provided by texts translated into English from the European national language of each member country, and that could cause some interference with the source language (Toury 1995) in Eurostat Yearbook texts.

2.6 ENSY corpus

The corpus was mainly explored by using AntConc 3.2.1. for listing keywords, and counting words and clusters. On the other hand the Corpus Query Processor - CQP (Christ 1994) was used to study sentence length and parts of speech to compare their use in the subcorpora. For this purpose ENSY parts of speech were tagged at the University of Bologna[34].

Each subcorpus was tagged separately in order to enable comparisons according to the following combinations:

1) *Natstat* vs. *Transtat* English;

[34] Dipartimento di Scuola Superiore di Lingua Moderne per Interpreti e Traduttori (SSLMIT), Forlì – Italy.

2) *Eustat* + *Transtat* vs. *Natstat* English;

3) *Eustat* + *Natstat* vs. *Transtat.*

The first combination was used to compare native English language (i.e. texts drafted in English by native English speakers and addressed to native English speakers) with translated English language (i.e. texts drafted in a language different from English, afterwards translated by native English speakers, and addressed to an international public of native/non-native English speakers).

The second combination was used to compare EU English language (i.e. texts written by native English speakers operating in an international context and addressing an international public of native/non-native English speakers) together with translated English language (i.e. texts drafted in a language different from English, afterwards translated by native English speakers, and addressed to an international public of native/non-native English speakers). Therefore both *Eustat* and *Transtat* have internationality in common. That is why *Eustat* and *Transtat* are associated and compared to *Natstat* which collects native English language texts (i.e. texts drafted in English by native English speakers and addressed to native English speakers).

The third combination was used to compare texts originally drafted in English (*Eustat* and *Natstat*) with translated texts (*Transtat*).

The first step of the investigation was to extract the keyword list of the three above-mentioned cross-comparisons and to examine occurrences of both function and content words and their combination in clusters.

All reported examples on words and clusters extracted from the corpus were tested to avoid the presentation of single occurrences. This means

that those patterns occur at least more than once in the corpus or subcorpora. Anyhow, whenever occurrences were in terms of units, this was remarked in their presentation. When analyzing occurrences and features from *Transtat* a check was made to ensure that features, patterns or words occurred in more than 7 National Statistical Yearbooks from 7 different countries. In this way there is evidence that a specific feature is not related to a specific source language.

In particular, the number of nouns and adjectives was compared to the whole number of words both in ENSY and in the three subcorpora to assess the use of nominalization (cf. Gotti 2011; Taviano 2010). The result was then compared to the percentage of verbs over the number of words to estimate the relevance of this pattern.

Lexical density (cf. Gotti 2006; Laviosa 2002; Baker 2001; Oholan 2004) was measured by means of the ratio between lexical and non-lexical words. The former group included: adjectives, nouns, non-auxiliary verbs and adverbs (-ly); the latter included: conjunctions, determiners, prepositions, modals, personal / possessive pronouns / adjectives, particle, *wh*-determiners and pronouns (i.e. ‘who’, ‘what’), non-lexical adverbs (e.g. ‘where’, ‘when’, ‘more’, ‘most’), auxiliary verbs, others (e.g. existential ‘there’, ‘such’, ‘quite’), numbers. The results from ENSY were taken into account as features of specialized discourse. This quantitative analysis was also carried out in the three subcorpora in order to compare ELF and native English.

This lexical density was related also to simplification as a universal feature of translated texts (Laviosa-Braithwaite 1996: 119), a lower lexical density being evidence of simplified texts.

The investigation of the ECC (English Comparable Corpus, see 2.4) focuses on global aspects of lexical and stylistic simplification and reveals four consistent patterns of lexical simplification in translated vs. original texts, independently of text category. These patterns are: relatively lower proportion of lexical words versus grammatical words; relatively higher proportion of high frequency versus low frequency words; relatively greater repetition of the most frequent words and less variety in the words most frequently used (cf. Laviosa 1996)[35].

Other 'Translation Universals', such as 'explicitation', were counted referring to some specific expressions: 'in order to', 'as well as', the use of determiners, anaphoric reference, and *of*-phrases preferred to '*s*-genitive.

Type/token ratio was measured on similar-sized files of the three subcorpora and then cross-checked in order to avoid size bias which is very high in such specialized texts (Biber 1999).

Average sentence length was also calculated as a feature related to specialized discourse (Gotti 2011: 65) and to translation universal features (Oholan 2004).

The *Longman Grammar of Spoken and Written English* (LGSWE) (Biber *et al.* 1999) was used as reference for standard language. This means that some features such as the use of the determiner 'the', 'may' and 'can', and 'lexical bundles' were compared with LGSWE findings, and in particular with the academic prose register to study their use in the discourse of statistics.

[35] http://www.llc.manchester.ac.uk/ctis/phd/completed_phd/laviosa/ (Last accessed 7 May 2012)

2.7 ENSY context

In order to draw a wider picture of the discourse of statistics, a survey was carried out to contextualize the texts included in the corpus. Statisticians prepare tables and write texts which comment upon statistics. They are the ones who write the National Statistical Yearbooks. Usually, each group of statisticians who have collected and processed data on a specific topic are the ones who write texts on that topic. For this reason, their way of dealing with statistical language is one of the aspects to be investigated. Since they are responsible for the dissemination of statistical knowledge, clarity and accessibility in statistics depend mainly on their choices and approach in describing data for dissemination.

2.7.1 Survey on language and statistics

The questionnaire for the survey was draft on the basis of *Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System*[36].

The questionnaire is titled "Informative Questionnaire on Language in Statistics"; it is given in English and is divided into two main sections. The first section on 'Respondent data' includes twelve questions on the respondent's background, English proficiency and use. The second section is titled 'Clarity and Accessibility' and, by means of thirteen questions, aims at testing how clarity and accessibility are relevant to statisticians, and perceived by expert users in statistical yearbooks. The second section is the most relevant in relation to the present study. The

36 http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPSQDET27062006.pdf (Last accessed 18 October 2011)

collected questionnaires inform us on how and if statisticians perceive clarity as a goal to be achieved. At the same time they are requested to express their perception of clarity when reading Eurostat and National Statistical Yearbooks. This type of information can be very useful when it is compared with the data collected by corpus-assisted analysis as findings empirically provide evidence that some of the limits noticed in the study are also perceived by that part of the scientific community most deeply involved in the drafting of statistical publications.

To be tested was prior to data collection, the questionnaire was submitted to five Italian statisticians. Some corrections were made after testing. However the questionnaire was not aimed at providing a comprehensive and ultimate response on interrelationships between statistics and language. It was an attempt to get to know how statisticians and other statistical experts deal with the discourse of statistics, and more specifically with the English language when publishing and reading data. Our ultimate aim, as mentioned, was to focus on some persisting difficulties and some achievements in the awareness that statisticians have of the need to improve clarity and accessibility in statistical texts.

The questionnaire contained no open questions that might facilitate comparisons and interpretations. Only multiple-choice questions were given and in two cases they included “other, specify...”. Those questions were related to suggestions to improve clarity and accessibility and blank space was purposely left for interviewees to write down different suggestions.

It was a limited survey on a small sample (43 questionnaires) and with a reduced burden for respondents. Filling-in the whole questionnaire required from 5 to 10 minutes. That characteristic facilitated responses

by a higher number of people and also by high-level and senior statisticians.

The questionnaire was sent by e-mail to all European National Statistical Institutes of EU member countries (EU-27) and Eurostat, asking them to submit it to the people in charge of the Yearbook drafting. Thirteen NSIs did not reply at all. UK Statistics replied that, since they were not publishing the National Statistical Yearbook any longer, they would not participate in the survey. Some other NSIs replied through the Public Relations Bureau, whose representatives filled-in the questionnaire.

The total number of filled-in questionnaire from NSIs and Eurostat is forty-three, and they are from: Bulgaria, Czech Republic, Cyprus, Estonia, Finland, Italy, Lithuania, Romania, Slovenia, Switzerland, Sweden, Hungary, Ireland and Eurostat. Switzerland is included not as a EU member country, but as a member of EFTA (European Free Trade Association), which has a full cooperation (except voting) in Eurostat matters.

The questionnaire could be considered a 'total survey', since all EU National Statistical Institutes were contacted as the entire population of the survey, but the 43 filled-in questionnaires are from 14 countries, and therefore 'no-response' due to 'unreturned questionnaire'[37] was to be analyzed. This high 'no-response' rate could be justified by different publication policies within EU NSIs, as explained above (see section 2.2).

The majority of returned questionnaires are from countries whose yearbooks are included in this study. This highlights a different approach on the part of NSIs to data dissemination. Those NSIs who do not

[37] For a statistical definition of 'No-response' see: http://stats.oecd.org/glossary/detail.asp?ID=3764 (Last accessed 19 October 2011)

publish the English translation of their publications online might not be interested in participating in a survey on this topic. As already described, in spite of EU recommendations, some countries have not increased the number of English translated publications; at times they have increased the number of on-line data to enhance accessibility, but they are still presented in the national language and therefore cannot be reached out by the international public. The following Table reports information useful to compare questionnaire responses and online publications in English. The third column reports information about whether the National Statistical Yearbook in English is available online:

COUNTRY	N. OF FILLED-IN QUESTIONNAIRE	ENGLISH YEARBOOK ONLINE
Austria	0	
Belgium	0	
Bulgaria	1	YES
Cyprus	1	YES
Denmark	0	YES
Estonia	3	YES
Finland	1	YES
France	0	
Germany	0	
Greece	0	
Ireland	2	YES
Italy	20	YES
Latvia	0	

Lithuania	1	YES
Luxembourg	0	
Malta	0	
Netherlands	0	YES
Poland	0	YES
Portugal	0	YES
United Kingdom	0	
Czech Rep.	5	YES
Romania	2	
Slovakia	0	YES
Slovenia	1	YES
Spain	0	
Sweden	1	
Hungary	1	YES
EUROSTAT	3	YES
Switzerland	1	YES
TOTAL	43	

Table 4 – Comparison between the publication of National Statistical Yearbook in English and participation in the Survey on Language in Statistics.

As shown in Table 4, all filled-in questionnaires were sent back by NSIs whose National Statistical Yearbook in English is available online. As for Sweden and Romania, it should be noticed that even though their Statistical Yearbooks in English are not available online their official websites present several statistical publications completely in English. We suppose that this is the reason why they were interested in

responding to the survey, hence they could be assimilated to the others as far as concerns accessibility by the international community.

All questionnaires were submitted and filled-in by the end of June 2011.

3. DATA ANALYSIS AND FINDINGS

This chapter is divided into subsections which analyse discursive features in statistics texts. Each sub-section provides information on occurrences and patterns; more specific findings are illustrated by means of subcorpora cross-checks with a view to highlighting differences which characterize *Transtat, Eustat* and *Natstat.*

In some cases, LSWE corpus (Longman Spoken and Written English Corpus 1999) was used as a standard reference to emphasise differences and similarities with ENSY. The LSWE corpus is characterised by four different registers, namely conversation, fiction, news and academic prose. Among these registers, academic prose was considered as a main reference, as it includes research articles and book extracts from a wide range of academic disciplines, including sciences, social sciences, and humanities, and hence can be considered more homogenous in relation to the texts under examination.

In the following sections, many examples are reported from the three subcorpora as evidence of the analysis that has been carried out, to highlight a specific usage and make the reader familiar with the discourse of statistics.

All retrieved data are accompanied by figures and tables which provide a visual representation to the descriptions.

Some of the data collected proof ELF-related features which differ from the native language. In such cases, the comparison focuses on differences and similarities between *Eustat* and *Transtat. Eustat* makes use of the language of native and non-native English speakers communicating in an international setting (i.e. the EU); on the other hand *Transtat* presents

translations into English addressed to an international public of native and non-native English speakers. The international use of both subcorpora is evidenced by some shared patterns which are worthy investigating.

Specific features of translated yearbooks are highlighted with reference to Translation Studies by means of data extracted from *Transtat* and compared both to *Eustat* and *Natstat.* The specific feature of translated English can be analyzed thanks to *Transtat* which collects translations from different source languages, and analyses similar features among them: the so-called Translation Universals (Mauranen / Kujamäki 2004).

Specialized discourse features have been found to be common to the three subcorpora but to a different extent. Specific characteristics of each subcorpus have been analyzed to describe their peculiarities.

3.1 ELF and the discourse of statistics

3.1.1 Americanization and colloquialization

As reported in section 1.5, ELF is characterized by some features which show preference for American English over British English. This trend was also identified as colloquialization. This is a significant stylistic shift in the twentieth-century English in which the written norm moves towards a reduction of differences with the spoken language, and towards greater tolerance of informality (Hundt / Mair 1999).

Some of these patterns were investigated in ENSY and are discussed in the following section.

3.1.1.1 ‘-ised’ and ‘-ized’ spelling

The first and most known feature of American English is the difference in spelling ‘s’ and ‘z’, hence words with ‘s’ and ‘z’ endings were searched in the corpus. The occurrences of ‘-isation’ as opposed to ‘-ization’, and ‘-ised’ as opposed to ‘-ized’ were counted. The result was that the ‘z’ spelling has a relevant number of occurrences in *Transtat* only, and expectedly no occurrences in *Natstat*. The frequency of occurrences is reported in Figure 3:

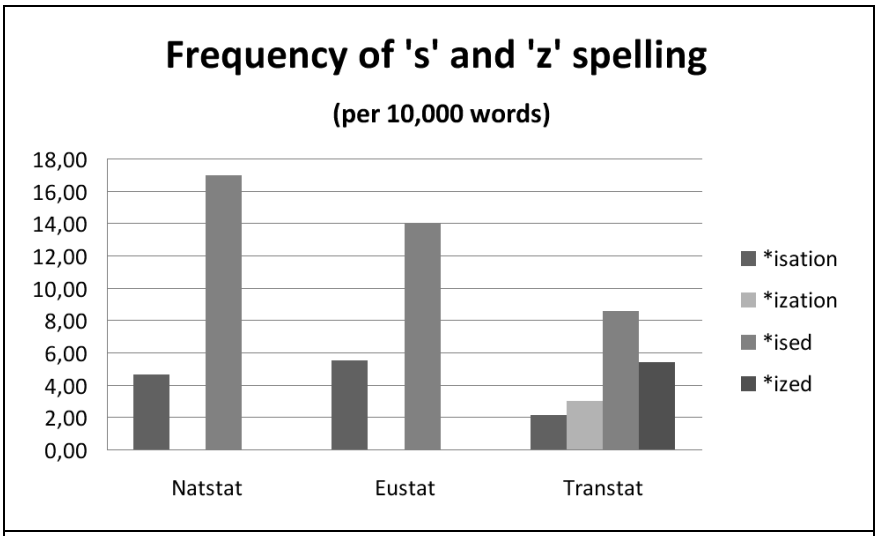

Figure 3. Frequency of ‘s’ and ‘z’ spelling per subcorpora.

It could be assumed that the occurrences of ‘z’ spelling in *Transtat* were due to translations made by American native speakers. However, by manually checking the corpus both ‘s’ and ‘z’ were found within the same files. Therefore ‘-ization’ and ‘-ized’ were included in ELF features and cannot be attributed to the origin of the translator. It was also observed that the American ending is not used in *Eustat*, which prefers the British spelling. Three occurrences of ‘-ization’ were found in Eustat

but only for World Health Organization (WHO), hence it does not appear a decision of the drafter.

Another reference to ‘z’ or ‘s’ spelling is on nominalization. Taviano (2010: 9) notices that the use of nominalization is less extensive in British texts. Figure 3 above provides evidence that ‘-isation’ and ‘-ization’ have a very similar number of occurrences per 10,000 words in *Eustat* (5.6 per 10,000 tokens) and *Transtat* (5.1 per 10,000 tokens), which is higher than in *Natstat* (4.6 per 10,000 tokens), while ‘-ised’ and ‘-ized’ rank higher in *Natstat* (17 occurrences per 10,000 tokens), and lower in *Eustat* (14 occurrences) and *Transtat* (14 occurrences). Verb forms ending in ‘-ised’ are more used than nouns ending in ‘-isation’ in native English yearbooks, which confirms the preference of verbs over nouns. The use of verbs instead of nouns is also recommended by *Clear Writing* – tip 6 – (see 1.7), but still nominalization is recognised as an ELF feature (Taviano 2010: 27) and characterises both *Eustat* and *Transtat* with equal numbers in percentage terms, at least with reference to nouns ending in ‘-isation’ and ‘-ization’.

3.1.1.2 Semi-modals

In ENSY, semi-modals are not very frequent. Nonetheless, they were studied in order to investigate another feature of colloquialization and Americanization with special reference to *Eustat* and the use of ELF. Semi-modals are now being used in written texts along with *s*-genitive as forms of spoken language (Krubg 2000; Hundt / Mair 1999; Hinrichs / Szmrecsanyi 2007).

The semi-modal forms listed by Biber *et al.* (1999: 484) were searched in the corpus as results from Table 5:

Occurrences of semi-modals						
	Natstat	**F.p.m.w.**	**Eustat**	**F.p.m.w.**	**Transtat**	**F.p.m.w.**
Have/has to	***0***		***39***	***117.35***	***72***	***98.58***
Had better	0		0		0	
Got to	0		0		0	
Supposed to	0		0		0	
Be going to	***0***		***0***		***2***	***2.74***
Need to	***0***		***35***	***105.32***	***6***	***8.22***
Ought to	0		0		0	
Dare to	0		0		0	

Table 5- Occurrences of semi-modals[38] per subcorpora.

The Table reports the low frequencies of semi-modals, which are completely absent in *Natstat.* Only three forms of semi-modals were found in ENSY: 'have/has to', 'be going to' and 'need to'. *Eustat* has the highest frequency per million words (222.67), which is double to *Transtat* frequency (109.54). Even though with a low frequency, semi-modals confirm the trend towards colloquialization for *Eustat.* For some specific features (i.e. 'have/has to', 'be going to', 'need to') this trend is also present in *Transtat*, although more numerous occurrences are found only in some translations, namely Estonia and The Nederland's files. Thus, this feature cannot be considered typical of *Transtat.* However, it is interesting to notice that some forms of colloquialization are used in translations.

3.1.1.3 *S*-genitive

Data related to the use of *s*-genitive were searched to further investigate colloquialization in ENSY. Nonetheless the use of *s*-genitive compete

with that of *of*- clauses, and this aspect will be deepened later in this chapter. Not all *s*-genitive occurrences are a possible alternative to *of*-clauses (i.e. fixed *of*-expressions, *of* preceded by a verb, etc.). Nonetheless, they can provide some useful information in the comparison among the three subcorpora on the use of more explicit (*of*-phrases) and implicit (*s*-genitive) features.

Generally speaking, this feature is not very frequent in the discourse of statistics. This becomes evident in the comparison with LGSWE academic prose. Biber *et al.* (1999: 301) remark that the latter has a "surprisingly" low frequency of *s*-genitive (2,500 occurrences per million words); in the present study frequency is even lower.

S-genitive in *Natstat* is used in one single sentence repeated four times:

(8) […] which measures Central **Government's** net surplus or borrowing position. (Natstat)[39]

In *Transtat*, the *s*-genitive structure is used in 6 translations only, and 71% of occurrences are concentrated in Denmark files. For these reasons this pattern cannot be considered relevant in *Transtat* (see section 2.6). *Eustat*, which has the highest relative frequency (1,700 occurrences per million words), shows a preference for this pattern in the place of *of*-phrases.

The more frequent use of *s*-genitive in *Eustat* could explain at least in part why 'of' has a high keyness (it ranks 12) when *Transtat* is compared with *Eustat*.

[38] In Table 5 columns named Natstat, Eustat and Transtat report number of occurrences; in 'F.p.m.w.' columns, frequency is normalized per million words for each subcorpus.

[39] Repetition of the same sentences is due to the process of cut and paste which characterizes Statistical Yearbooks.

It is worth noticing that the *s*-genitive has a very peculiar use in *Eustat*, as is evidenced by examples (9)-(12). In the majority of cases (67%), the *s*-genitive occurs in relation to the EU, Eurostat and Europe:

(9) **Eurostat's** transport statistics describe the most important features of transport […] (Eustat)[40]

(10) The **EU's** population is characterised by a relatively high life expectancy at birth […] (Eustat)

(11) About 13% of the **EU-25's** territory was considered as a protected area […] (Eustat)

(12) The maps presented here illustrated the diversity of **Europe's** regions**.** (Eustat)

The use of *s*-genitive in the place of *of*-phrases is very rich, and has been widely investigated especially with reference to press/news language (Biber 1999, 2003; Hinrichs / Szmrecsanyi 2007). In the present case various factors concur to this feature. A general interpretation is provided by Biber (1999: 307):

> There is another subject-like characteristic of the nouns which tend to appear in *s*-genitive: they refer to entities which are likely to be found as themes in texts.

This could explain why nouns referring to the EU are very frequently used with the *s*-genitive; an additional interpretation is that in a more informal setting there is a greater preference for the *s*-genitive (Altenberg 1982). This can help us to understand how Eurostat addresses its readers. The use of patterns which are borrowed from spoken language make the register more informal and the text closer to its readers. We can also

[40] Bold added in all examples for emphasis.

recognise a feature of colloquialization, a tendency towards "partial rapprochement" between spoken and written norms (Hundt / Mair 1999), which has been also described as "democratization" of the written norm (Fairclough 1992).

Altenberg (1982) and Biber (1999) recall that animate possessors prefer *s*-genitive, whereas inanimate possessors prefer *of*-genitive, and this is the tradition of prescriptive grammars. Eventually this use of the *s*-genitive with EU entities can be also interpreted as a sort of personification / animation of European bodies that emphasises the human aspect vs. bureaucracy. As noticed by Piga (2011), "The desire to reduce social distance and promote a tangible sense of proximity with EU citizens […] can do more to 'give a human face' to the information they provide".

3.1.1.4 'Can' and 'may'

Another feature shared by *Eustat* and *Transtat* is the type of modals they use. Modal occurrences were counted in the three subcorpora, and the four top-frequent modals in ENSY are presented with their distribution in Figure 4:

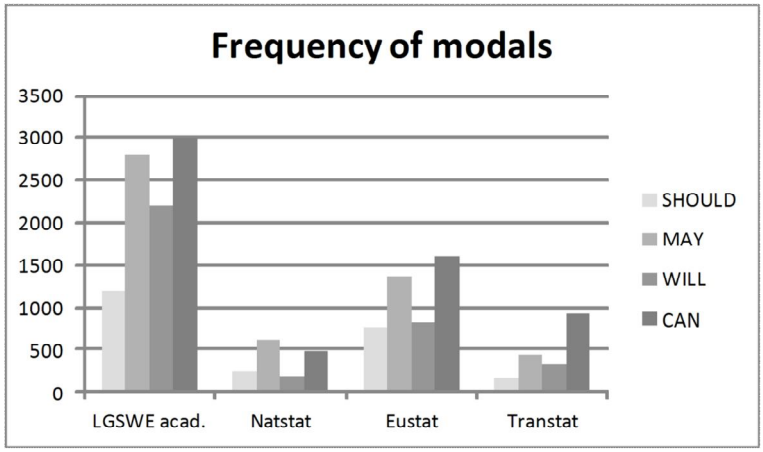

Figure 4. Frequency of 'should', 'may', 'will' and 'can' (per million words).

Eustat and *Transtat* appear to have a very different frequency of modals but the same trend (i.e. 'can' is the most frequent, 'may' ranks second, followed by 'will' and 'should'). *Natstat*, instead, has the lowest frequency of modals, and a different trend (i.e. 'may' ranks first, followed by 'can', 'should' and 'will' is the order). Hence, in the three subcorpora, 'can' and 'may' are the most frequent and this is a feature they share with LGSWE (academic prose).

'Can', and 'may', were thus studied as the most relevant modals in ENSY. There are two different traditions in the semantic analysis of 'can' and 'may', as for modals in general. One tradition attributes to each modal one unified meaning, 'monosemy', which is differently interpreted with reference to context (see Perkins 1983; Groefsema 1995; Klinge 1993 among others). Conversely, the tradition of 'polysemy' views each of the modals as polysemous, i.e. expressing two or more independent meanings (see Palmer, 1990; Coates, 1983 among others). In this research, we refer to the 'polysemy' approach to analyse modality and in particular to a binary scheme which distinguishes between two different types of meanings typically referred to as 'deontic' and 'epistemic' (Collins 2006/2009, Biber 1999). Biber (1999: 491) labels the meanings of modals as 'intrinsic' and 'extrinsic' respectively, and points out that deontic / intrinsic modals usually refer to actions directly controlled by an agent, possibly human beings, with a dynamic main verb. The deontic meanings express permission, obligation or intention; the epistemic meanings express possibility, prediction and necessity and the main verb is usually static. The two most recurrent modals in ENSY have, following Biber *et al.* (1999: 485), the meanings listed in Table 6:

MEANINGS		
	Epistemic/extrinsic	**Deontic/intrinsic**
May	Possibility	Permission
Can	Possibility	Ability

Table 6- Meanings of modals according to the polysemy theory.

All the occurrences of 'may' and 'can' were checked manually in the corpus in order to analyse the presence of human and non-human agents. It was found that the presence of human agents for modals is extremely rare, and these few cases are shown in the examples below:

(13) A claimant **can** however also claim income-based Jobseekers Allowance which is not included in the past Unemployed Benefit comparative figures. (Natstat) <possibility>

(14) One person **can** have more than one job. (Natstat) <possibility>

(15) Note that one person **may** have more than one subscription. (Eustat) <possibility>

(16) A Japanese child born in 2008 **can** expect to reach the age of 82. (Transtat) <possibility>

It has to be noticed that in examples (13), (14), (15) and (16) the epistemic meaning prevails even with human agents in statistics discourse, and this occurs in the whole corpus. The reported examples typically refer to information provided to the readers in order to enable them to interpret data, and why some categories are included or excluded from the figures presented. Deontic meaning is virtually not used and is found in methodological notes and introductions only. Even when *can* could be interpreted with a prevailing deontic meaning, epistemic interpretation is still possible as is shown in example (17):

(17) The energy intensity of an economy **can** be measured by the amount of energy consumed to produce one unit of GDP. (Eustat)
<ability> & <possibility>

The meaning of (17) can be either the explanation on the only method to measure energy intensity (deontic) or a possible way, among others, to measure energy intensity (epistemic). In this specific case the context doesn't enable us to understand whether it is the only way or one among a many.

The meaning of 'can' is developing, and the studies on the topic confirm our findings. Coates (1995: 63) points out that 'can' may be developing 'genuinely' epistemic uses; Collins (2007) notices an increased use of 'can' in its epistemic meaning and writes: concluding his article "C*an* and *may*: monosemy or polysemy?"

> […] there are signs that the epistemic possibility sense of *can* is becoming established (as we might expect to happen, historically), as it sheds its syntactic/semantic restriction to non-affirmative contexts.

In ENSY, some evidence was found of the interchangeable use of 'can' and 'may'. Collins (2009: 91) claims that:

> "the two modals of possibility *can* and *may*, share a high level of semantic overlap, so it is not surprising that there has been a good deal of attention paid to the relationship between them in the literature".

The semantic overlap in the discourse of statistics is evidenced in some recurrent expressions where *Transtat* can be noticed to prefer 'can' and *Natstat* 'may'. Some comparable examples showing the preference of one modal over the other in the two subcorpora are reported below:

(18) Figures in the same tables in different yearbooks referring to the same year **can differ** slightly from each other. (*Transtat*) <possibility>

(19) The data in Table 5.9 **may be slightly different** from data in last year table. (*Natstat*) <possibility>

(20) In some cases sums of components **can differ** from the amount given in the item. (*Transtat*) <possibility>

(21) Collection methodology **may differ**. (*Natstat*) <possibility>

This use of 'may' and 'can' which can be interchangeable, can be attributed to the changes happening in language in the last decades which are leading to the decline of 'may' especially in American English, and less in British English (Leech 2003). This process provides the explanation for the high ranking of *may* in *Natstat* differently from *Eustat*, *Transtat* and LGSWE academic prose, which are more influenced by American English. *Eustat* confirms features of colloquialization which are also considered features of Americanization (Hinrichs / Szmrecsanyi 2007: 442), and this has been already noticed also for other aspects (see 3.1.1.2). In this study these features can be related to the use of ELF in EU contexts. As for *Transtat*, the use of 'can' is preferred for the same reasons exposed for *Eustat*. Granger (2008) notices that also learners prefer the use of 'can'. In the case of LGSWE academic prose, texts included in the corpus are also American and this justifies the higher frequency of 'can'. Both 'may' and 'can' mark logical possibility and this is how they are used also in ENSY. The trend to prefer 'can' to 'may' could be included in those features of Americanization which involve 'Eurospeak' both when drafting English texts for the EU and

when translating texts from European languages into English to reach out an international public.

3.1.2 Nominalization in ELF

In many of the studies on ELF written texts (Murphy 2008; Taviano 2010; Jenkins 2006 among others) nominalization is identified as a typical feature of ELF. As to the discourse of statistics, this is influenced by nominalization as a feature of specialized discourse (Gotti 2006). The occurrences of 'the *ion of' in the three subcorpora were counted as evidence of this feature. An example of the pattern 'the *ion of' is reported in example (22) below:

(22) Environmental requirements and standards have been set and financial instruments implemented for **the** reduct**ion of** negative environmental impacts of the growing production and consumption. (*Transtat*)

(23) At a European level, statistics are increasingly important for **the** definit**ion**, implementat**ion**, monitoring and evaluat**ion** of policies. (*Eustat*)

In (22) 'the reduction of' could be easily replaced by the use of the verbal form 'for reducing'; in (23) instead of 'definition, implementation and evaluation' verbal forms could have been used as in the case of 'monitoring'. Figure 5 below highlights differences of occurrences '*ion of'.

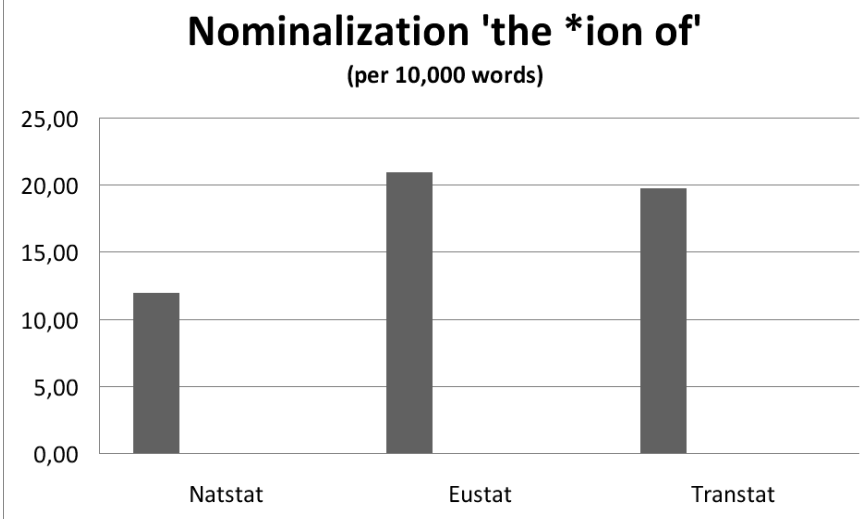

Figure 5. Occurrences of the pattern 'the *ion of' in the three subcorpora.

The highest frequency of words with this ending was detected with similar numbers in *Eustat* and *Transtat*; *Natstat,* instead, shows a less frequent use of such words confirming the difference in the use of nominalization between Native English and ELF.

3.2 Translation features

Transtat was compared to the other subcorpora in order to find out features related to translation. To this aim the keywordlist of *Transtat* referred to *Eustat* + *Natsat* was extracted and some high-ranked function and content words were chosen for analysis. Also some clusters were examined as typical of *Transtat.*

3.2.1 Anaphoric reference

Biber *et al.* (1999: 263) writes: "The definite article […] specifies that the referent of the noun phrase is assumed to be known to the speaker

and the addressee, in which case we speak of anaphoric reference". In our statistical corpus, the definite article 'the' has a very high frequency. In LSWE, the distribution of the definite article across registers results to be the highest in academic prose, where it reaches about 55 thousand occurrences per million words. In ENSY (see 2.5), instead, the frequency is much higher than in LGSWE. The pick is in *Transtat* as shown in Figure 6:

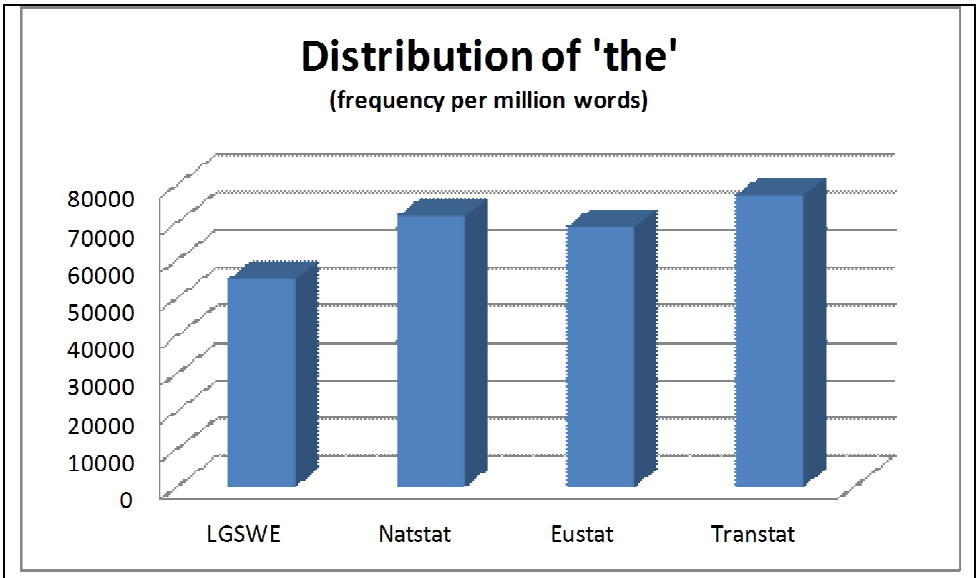

Figure 6 – Distribution of the occurrences of 'the' in LGSWE and the three subcorpora

The keyness of 'the' in *Transtat* was also investigated extracting keywordlists in cross-checks (Bernardini / Zanettin 2004). In particular, *Transtat* was compared to *Natstat* and *Eustat* separately. 'The' in the keyword list of *Transtat* compared to *Natstat* ranks 9, and 4 when compared to *Eustat*. The situation is completely different when *Eustat* is compared with eiher *Natstat* or *Transtat*. In those cases, 'the' is not ranked in the keywords (1-100); the same happens when *Natstat* is compared with *Transtat* and *Eustat*. Keywordlist investigation confirms

that the use of ‘the’ is a specific feature of *Transtat*, which reinforces this general characteristic of the discourse of statistics.

Literature on explicitation in translation considers ‘the’ as one of the strategies implemented to provide an explicit reference (Oholan 2004). Its wide use in *Transtat* cannot be considered language-systemic, which means attributable to interference by a specific language, since texts included in the subcorpus are translations from 13 different languages and not from one single language. All the translations give evidence of this high frequency. Hence, the use of ‘the’ should be interpreted as a means to specify references (Vanderauwera 1985) to the aim of grammatical explicitation. Also Murphy (2008), when comparing European edited and non-edited texts, refers to ‘the’ as one of the occurrences eliminated by editors, since a “freer” use of the definite article is also listed among characteristics of ELF (Seidlhofer 2004). As already noticed above, also in this case some features used by translators to the purpose of explicitation or disambiguation are also detected as ELF features.

Another pattern which could be associated to grammatical explicitation by means of anaphoric reference is also the use of ‘that/those + *of*-phrases’, whose distribution can be seen in Figure 7:

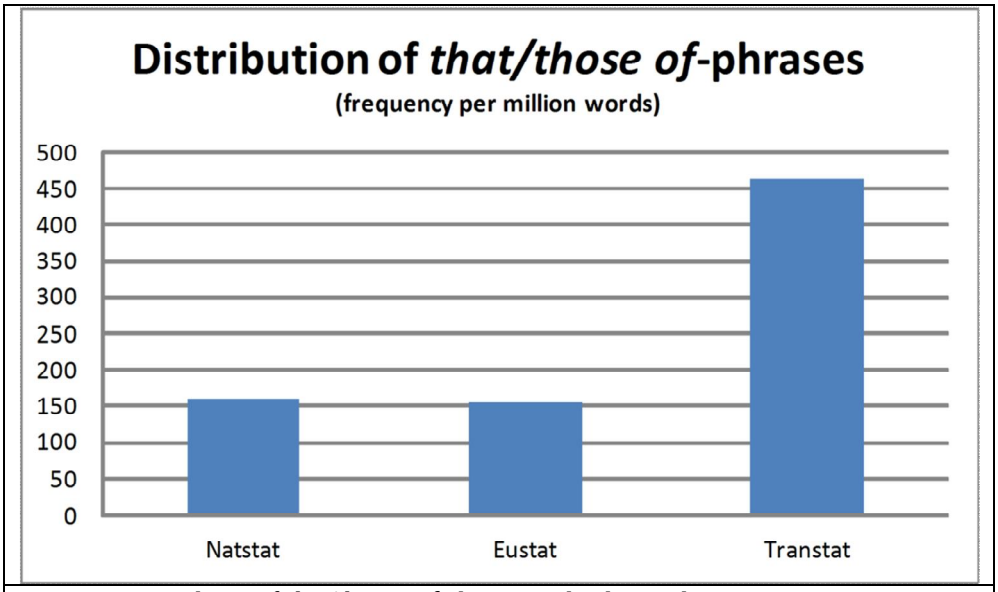

Figure 7 – Distribtion of *that/those +of*-phrases in the three subcorpora

This pattern is rare in *Natstat* and *Eustat*, and is more frequent in *Transtat.* That is why it can be included among the more general features of anaphoric reference which aims at making texts more comprehensible to readers. In statistics, this construction is used for comparisons in time, space and categories. Some examples are:

(24) In 2009 the number of administrative cases settled increased by 9, **that of** civil cases by 27 per cent. (*Transtat*)

(25) […] annual average turnover greater than **that of** retail enterprises. (*Transtat*)

In examples (24) and (25) the reference is very clear and explicit. The use of 'those of' and 'that of' provides the reader with univocal meaning. This is not the case of *Eustat* and *Natstat* which implement different strategies to maintain cohesion:

(26) This figure was 1.6 million higher than**(*)** the next largest student population, in the United Kingdom, and 2.0 million higher than **(*)** in France. (*Eustat*)

(27) Finland reported the highest proportion of R & D personnel (3.0%) as a share of the total labour force, with more than twice the EU-27 average **(*)** […] (*Eustat*)

(28) The average age of unmarried mothers is lower than **(*)** for married mothers. (*Natstat*)

(29) For pre-school going children the average hourly rate for paid childcare was highest in the Dublin region, at 5.15 over 24% higher than**(*)** the state average. (*Natstat*)

Examples (26) - (29) above show ellipsis of anaphoric reference graphically represented by (*) and no occurrences of *that/those of*-phrases in *Eustat* and *Natstat*, while cohesion is maintained by contextual features (Halliday / Hasan 1976) which avoid ambiguity, therefore reading examples (26) to (29) we have no doubts on the meaning. On the contrary *that/those of*-phrases concur to the aim of explicitation and, even more, disambiguation in *Transtat*.

3.2.2 Prepositions

Other high-ranked function words were studied to compare different features in the three sub-corpora and the result is a more frequent use of those words in translated texts.

High rank in the keyness of the function word ‘of’ can be observed in *Transtat* with reference to *Natstat* + *Eustat*: ‘of’ ranks 10, and maintains high keyness even when *Transtat* is compared to *Eustat* (12 rank) and *Natstat* (28 rank) separately. As claimed by Sinclair (1991), “[…] *of* is […] over two per cent of all the words, regardless of the kind of text

involved". In this specific case, the difference we want to emphasize is the keyness of this function word.

As to 'in', when comparing *Transtat* to *Eustat*, it ranks 10 in the keywordlist and gains a position when *Transtat* is compared to *Eustat* + *Natstat*; in this case, 'in' ranks 9.

An investigation 'of' and 'in' is worth while also because they are associated to 'the' by Murphy (2008: 55) in the list of occurrences eliminated by editors in EU English texts. The above exposed keyness of 'of' and 'in' is interesting as a specific feature of translated texts. Furthermore 'of' and 'in' concur both in "grammatical explicitation" and "disambiguation" (Baker 1996). Their use is specifically related to nouns that they introduce. Therefore, they will be analysed in the following section also in relationship with their collocate content words.

3.2.3 Noun repetition

Some content words characterise the discourse of statistics. The description of data and their analysis make use of very repetitive patterns and words that can be recognised in the corpus. In the following section we shall present some findings which show the language differences among the three subcorpora with respect to the most used nouns. The words were chosen by means of the keywordlist of *Transtat* with reference to the other subcorpora.

3.2.3.1 'Year/s'

As already explained (see section 2.2), data presented in statistical yearbooks refer to a specific time-frame, often a year. The word 'year' ranks second in the keywordlist of *Transtat* when referred to *Eustat*, and

third when referred to *Eustat* + *Natstat*. The word ‘year’ is relevant in all statistical texts, therefore its keyness in *Transtat* may be attributed to a higher number of occurrences and hence repetition of the word. For this reason its occurrences and collocates were studied in order to investigate differences across subcorpora:

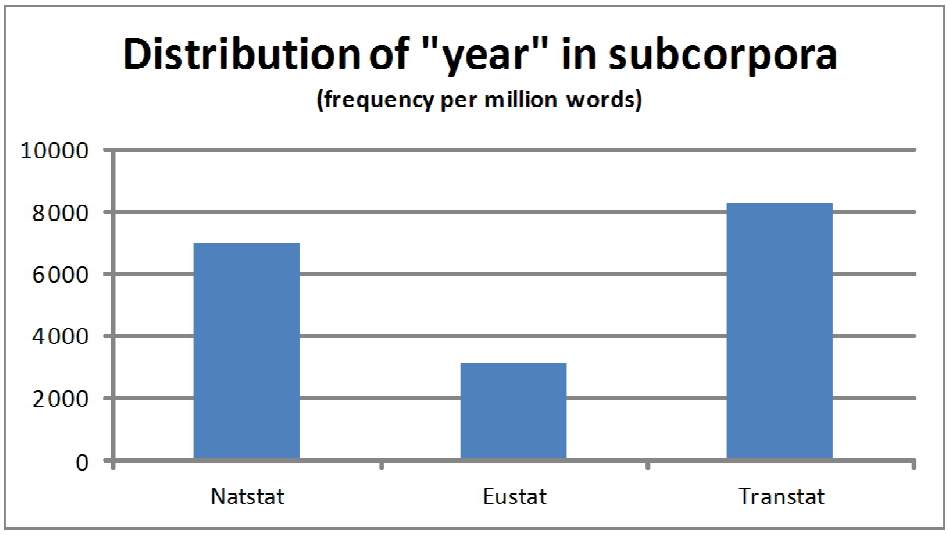

Figure 8 - Distribution of ‘year’ in subcorpora.

The most frequent left collocate of ‘year(s)’ is the determiner ‘the’ in all subcorpora, even-though with different percentages (34% in *Natstat*, 20% in *Eustat* and 42% in *Transtat*); these data confirm the findings in section 3.2.1 on the higher frequency of ‘the’ in *Transtat*.

Collocates of ‘year(s)’ greatly differ across the subcorpora as well as the related clusters do. For example, in *Transtat*, ‘previous’ is the most frequent left collocate of ‘year(s)’ after ‘the’, and recurs 925 times in this collocation. The same cluster has 21 occurrences in *Natstat* and only 10 occurrences in *Eustat*. Here are two examples on the use of this cluster:

(30) In 2007 the gallery Louisiana accounted for the highest admission rates of […] or 33% of visitors compared to the **previous year**. (*Transtat*)

(31) The economic growth that had been relatively high still in 2007 was the main reason why the labour market indicators were good compared to the **previous years**. (*Transtat*)

From the analysis of the three subcorpora, *Eustat* and *Natstat* were found to prefer the use of the year number instead of the phrase 'previous year' when comparing data to the year before, as in the following examples:

(32) States recorded a reduced deficit or increased surplus relative to GDP in **2008** compared to **2007.** (*Eustat*)

(33) This latest annual figure (**2008)** represented a reduction of just 0.1 percentage points in comparison with **2007**. (*Eustat*)

(34) Overseas visits to Ireland fell by 11.6% to 6,927,000 in **2009** compared to **2008.** (*Natstat*)

The use of 'previous' instead of numerals should be regarded as 'lexical explicitation' (Oholan 2004) to overcome ambiguities. The word 'previous' cannot be misinterpreted and refers to one-year-time span. On the contrary, when the numeral is reported, the reader is unconsciously obliged to count in order to be sure of time-span reference.

The use in *Natstat* and *Eustat* of year numbers is also interpreted as one of the concurrent reasons why 'year' is less frequent in *Eustat* and especially in *Natstat*, because year numerals are used with no explicit reference to 'year' - examples (32) to (34).

3.2.3.2 ‘Data’

A highly domain-specific word in statistical discourse is ‘data’. It is worth studying it as it is one of the most recurrent content words in the corpus and especially in *Transtat*. In *Natstat* the use of ‘data’ is less frequent and is replaced by near synonyms, such as ‘statistics’ or ‘figures’ as in examples (37) and (38) below. In these cases, we can speak of cohesion by means of ‘substitution’ (Halliday / Hasan 1976). In other cases an implicit lexical reference is used, as in examples (26) and (27). In (35) and (36) the reference to ‘data’ is omitted, and no pronoun or synonym is used but the context enables the reader to infer the meaning. This form of cohesion uses the device of ellipsis (Halliday / Hasan 1976).

(35) A break in continuity occurred in the Educational Attained series and, therefore, **data** for 2009 is not directly comparable with **previous years**. (*Natstat*)

(36) A variety of **data** relating to Northern Ireland are contained in the Appedix; caution should be exercised when comparing **these tables** with those of the Republic as collection methodology may differ. (*Natstat*)

(37) The source **data** on banking are principally drawn from the Central Bank […]. **The statistics** on public finance are obtained primarily from two administrative sources […] (*Natstat*)

(38) Historical **data** for the period up to and including 2006 are taken from the various censuses and registrations of deaths and births. **Detailed figures** for intercensal years are taken from the annual series of population and migration estimates. (*Natstat*)

The following Figure shows the higher frequency of ‘data’ in *Transtat* confirmed in occurrences normalised per million words. The main difference is in comparison with *Natstat*:

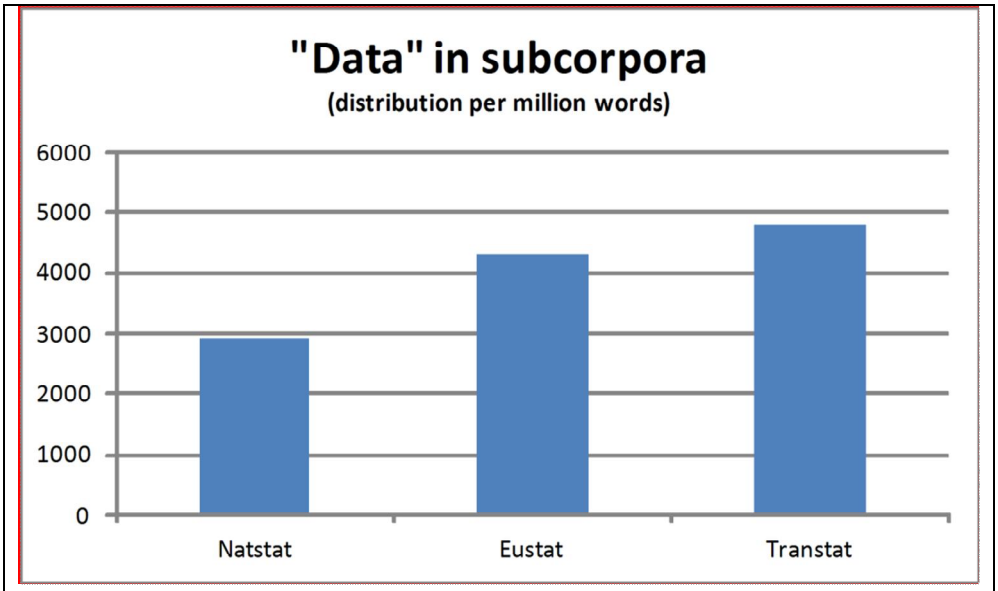

Figure 9 - Distribution of the word 'data' in the three subcorpora.

The first collocate of the word 'data' in ENSY is 'on', hence the cluster 'data on', was studied, and the result was quite a big gap between *Transtat* and *Natstat*+ *Eustat*. In *Transtat*, the right collocate of 'data' is 'on' in the percentage of 33.8%; the percentage decreases to 10.4% in the case of *Natstat* + *Eurostat*. When analysing *Natstat* separately, 'on' follows the word 'data' (12.8 times out of 100 occurrences). In the case of Eustat the percentage is very low, i.e. 10%.

The distribution of 'data on' per subcorpora can be added to features considered as a strategy of 'explicitation' implemented by translators. This trend was also confirmed by findings exposed above in the present chapter (see section 3.2.2). The use of 'on' to describe data avoids ambiguities and presents explicit references. The following groups of examples provide information on how the cluster 'data on' is used in *Transtat*; examples provide comparisons to similar expressions in *Eustat* and *Natstat* where different patterns are used:

(39) This section presents **data on the balance of payments.** (*Transtat*)

(40) **The balance of payments data** […] (*Natstat*)

(41) Ministry of Agriculture provides Statistics Estonia with **data on fish catch** (*Transtat*)

(42) **Fishery data** (*Natstat*)

(43) **The data on GDP** are not comparable […] (*Transtat*)

(44) This section presents **data on gross domestic product.** (*Transtat*)

(45) **GDP data** used as FISIM allocated […] (*Natstat*)

(46) The aim of the survey is to collect **data on the employment.** (*Transtat*)

(47) These additional statistics include **employment data.** (*Eustat*)

(48) Statistical **data on health** allows the analysis of health care services. (*Transtat*)

(49) Subdivisions are enumerated to provide **data on healthcare.** (*Transtat*)

(50) **Healthcare data** presented in this section […] (*Eustat*)

Natstat and *Eustat* prefer the use of nouns as premodifiers instead of longer *on*-phrases. However, they seem to be more concise but less explicit.

3.2.3.3 'Age'

Information on age is a typical feature of data description, used to identify and divide people under study, for instance when age classes are

used to divide the population and to extract ratios and indices in the statistical domain. Hence, such term occurs very frequently and in different ways within statistical texts. In ENSY, it was studied with reference to 'aged', the participial adjective from the verb 'to age', to investigate both explicitation and nominalisation phenomena. The interest for this comparison stemmed from the observation that 'aged' was less used in *Eustat* and *Transtat*, and in particular occurred with anomalous collocates in *Transtat*. The following examples illustrate the latter phenomenon:

(51) Each sixth man and each fourth woman were **aged** 60 and older. (*Transtat*)

(52) […] especially those **aged** 15-24 years old. (*Transtat*)

(53) Women **aged** from 15 to 54. (*Transtat*)

(54) Aging index is the ratio between the old population (**aged** 65 years and over) and the young population (**aged** 0-14 years). (*Transtat*)

In *Transtat*[41], as shown in the examples above, 'aged' is accompanied by other words specifically related to age, such as 'older' (51), 'years old' (52), 'years' (54), or reference to time-span. 'Aged' with the right collocate 'from', as in example (53), can be found in *Transtat* only, it never occurs in the other subcorpora. A sort of misuse of 'aged' can be noticed in *Transtat*. This kind of collocations could be worth investigating in other domains and European translations, and could be included in those features that shifted from being regarded as mistakes to be included in ELF features (for example: as it was for the drop of 's' in the third person - Jenkins 1996). At this stage, we can see in these

clusters a sort of additional (redundant) specifications provided by *Transtat* when using 'aged', and we interpreted them as forms of disambiguation in translated texts.

In order to provide information on the different use of 'aged' in *Natstat* and *Eustat*, here follow some examples:

(55) Number of employed persons **aged** 15 to 64 expressed as a percentage. (*Natstat*)

(56) Figures are as a percentage of people **aged** 65 and over. (*Natstat*)

(57) Just over four fifths of births were to mothers **aged** 25 to 39 in 2009. (*Natstat*)

(58) Poland and Belgium recorded the highest share of adults **aged** 18 to 59 living in jobless households. (*Eustat*)

(59) Households headed by a person **aged** under 30. (*Eustat*)

(60) Coverage refers to those **aged** 16 to 74. (*Eustat*)

The use of 'aged' enables the writer to shorten the sentence, as we can see in examples (55) to (60). This more implicit kind of pattern is well exploited in *Natstat* but much less in *Transtat* and *Eustat* where 'age' is preferred to it. A comparison on the use of 'age' and 'aged' in the three subcorpora is reported in the Figure 10 below:

[41] Examples from *Transtat* are all from different source languages.

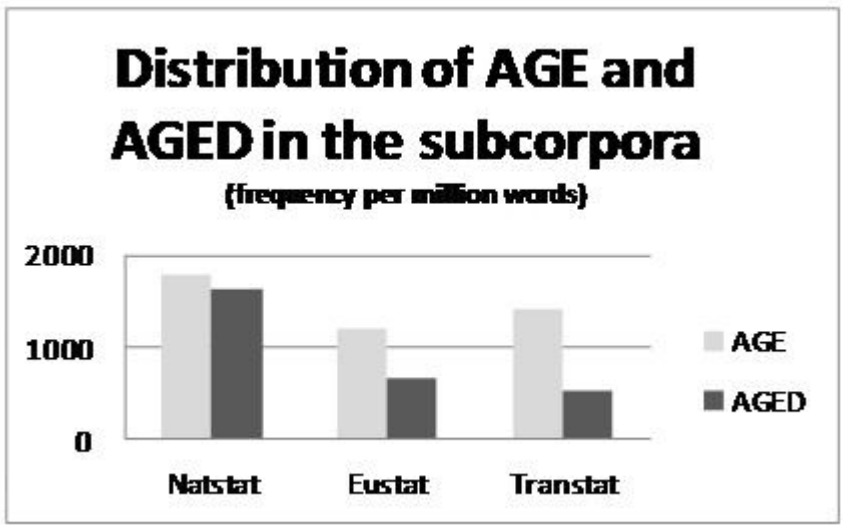

Figure 10 – Distrubution of 'age' and 'aged' in the subcorpora.

'Aged' is a word with one single meaning. It is very clear, that is why specifications are considered redundant. The preference for 'age' as a feature in translated and EU texts can be considered also as an example of nominalization, where the noun is used in place of a verb. As for *Transtat*, it can also be attributed to interference from the source language (Toury 1994) due to the fact that 'aged' has not a corresponding word in all languages, while 'age' has.

Here follow some examples on the use of 'age' which give evidence of its use in the three subcorpora:

(61) EU population in the 55-64 **age** group should be in employment […] (*Eustat*)

(62) Italy ranked as the country were people could expect to spend the longest period after the **age** of 65 in good health. (*Eustat*)

(63) The voters were citizens who had attained 18 years of **age** on the second day of election at the latest. (*Transtat*)

(64) Population in the **age** 15-64. (*Transtat*)

(65) Number of employed persons at the **age** of 15-64. (*Transtat*)

(66) Only a slightly greater proportion of females were in full time education at **age** 15. (*Natstat*)

(67) Male participation in the labour force for 15-19 **age** cohort. (*Natstat*)

(68) At the **age** of 20 […] (*Natstat*)

In *Natstat*, 'age' has 144 occurrences, and in 38 times only (26%) it is related to specific numerals signalling years of age. Some of these cases are reported in the examples (66), (67) and (68) above. In all other cases, 'age' is used to express concepts referred to age in general, and hence 'aged' could not be an alternative, as in the following examples:

(69) The increased population of working **age** […] (*Natstat*)

(70) Rates were evident across all **age** categories. (*Natstat*)

(71) The **age** of travellers […] (*Natstat*)

We can conclude that in *Natstat* 'aged' is the first option and the noun 'age' is used when necessary, while in *Eustat* and *Transtat* 'age' is used even when it could be replaced by the more concise form 'aged' as a tendency to nominalization. In *Transtat* the conciseness of 'aged' is not exploited and further specifications (cf. examples (51) to (54)) are added to explicitate its meaning, as a strategy implemented by translators (Oholan 2004).

3.3 Features of specialized discourse

In this section we shall investigate the discourse of statistics by means of features and measures which typically characterise specialized discourse. Therefore the use of the passive forms and personal/impersonal expressions occurring in ENSY are described together with the type/token ratio and the measure of average sentence length.

3.3.1 Use of the passive

Passive voice is considered among the typical syntactic features of specialized discourse (Gotti 2006). Evidence of its use was not found by means of keywordlist, but it was considered to be a relevant feature to describe statistics language within the family of a domain-specific discourse. As Gotti (2006: 74) remarks, “The pervasiveness of the passive may be accounted for by its usefulness as a depersonalising device in specialized discourse”. In our statistical corpus the use of the passive voice is quite frequent in all subcorpora, with some relevant differences which are shown in Figure 11:

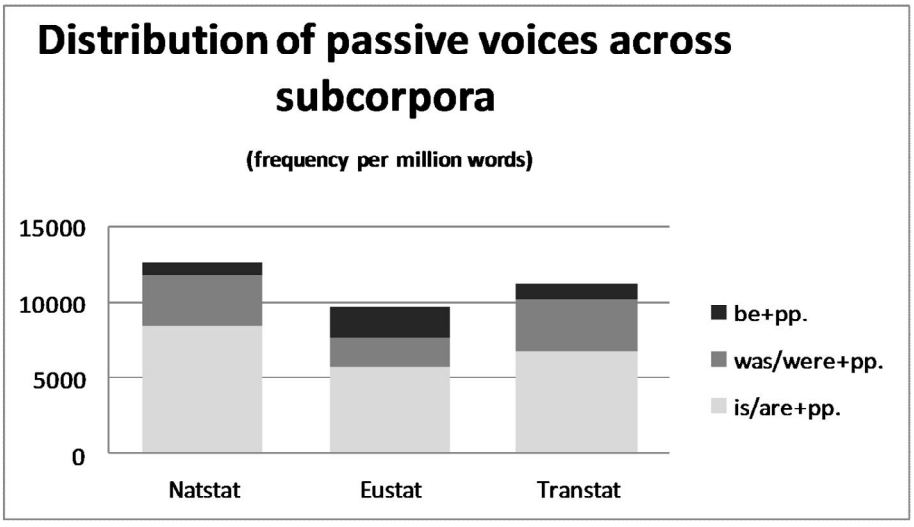

Figure 11- Distribution of passive forms in the three subcorpora.

Passive voice structures were found in the corpus searching for "be" inflections followed by a past participle; in the search also irregular verbs were included, checking all the list. Present tense is the most frequently used in all subcorpora ('is/are + past participle'), even-though it has the highest frequency in *Natstat*. This feature confirms the prevalence of Present Simple Passive with respect to all other passive tenses as also detected by Barber (1985: 8). Frequency of past tenses is the highes*t in Natstat*, followed by *Transtat*, and with a greater difference by *Eustat*. These data are in line with other features analyzed in the present study, where *Natstat* has the highest degree of conformity to specialized discourse features (see 1.3) and *Eustat* presents the most informal language of the three subcorpora (see 3.1.1). Comparing the subcorpora, *Natstat* has a highly frequent use of passive structures with 13,000 occurrences per million words, but has a lower frequency of the form "be+pp.", which is mostly used in *Eustat.* Some passive expressions like 'is/are given' are the most common in *Transtat*, 'is/are taken' in *Natstat*,

'may/can be found' in *Eustat*, where modals occur most frequently. Here follow some examples:

(72) Detailed figures for intercensal year **are taken** from the annual series of population and migration series. (*Natstat*)

(73) The point in time for the division into areas **is given** in the table heading or in the footnote. (*Transtat*)

(74) Large variations **can** also **be found** within a given country. (*Eustat*)

The hypothesis, not yet supported by data, is that the use of passive is avoided in the case of *Eustat* by means of verbs in the active voice whose subject is Europe and its bodies (EU, Eurostat etc.). This is in line with the tip of *Clear Writing* (2010) which suggests choosing the active rather than passive voice. The already mentioned personification of EU bodies (see 3.1.1.3) concurs to this feature. The following examples show this kind of expressions:

(75) Eurostat calculates the following ratio to compare 'rich' and 'poor'.(*Eustat*)

(76) Eurostat collects the data […] (*Eustat*)

(77) The EU adopted in 2005 the 'Integrated Guideline Package'. (*Eustat*)

(78) The European Commission assesses these programmes and the Council gives its opinion on them. (*Eustat*)

As for *Transtat,* the use of verbs in the active voice in the first plural person 'we' was detected as an alternative to passive forms. We will devote the next section to the analysis of this device.

3.3.2 The use of ‘we’

The pronoun ‘we’ is an instance of personalization and is also easier to use in English (Biber *et al.* 1999). It is hardly ever used in *Natstat* and *Eustat*, (2 and 8 occurrences respectively). The following examples report all the sentences in which ‘we’ is used in *Natstat* and *Eustat*; some sentences are repeated exactly the same two or more times in the same subcorpus:

(79) In recent years **we** have expanded the number of modules undertaken in any given year. (*Natstat*)

(80) It is also transforming the way in which **we** communicate, do business, and live everyday lives. (*Eustat*)

(81) The air **we** breathe contains gases. (*Eustat*)

(82) Rural development policy aims to ensure the survival of the countryside as **we** know it. (*Eustat*)

(83) The water **we** drink and bathe in are therefore major concerns all around the world. (*Eustat*)

(84) **We** depend on natural resources. (*Eustat*)

In *Transtat*, ‘we’ occurrences are 20 times more frequent than in the other subcorpora (i.e. 428 occurrences per million words). The pronoun ‘we’ occurs in all *Transtat* files but Cyprus and Italy’s. It has to be noticed that the numbers for ‘we’ are highly biased by its very frequent use in Slovenia which has almost 2000 ‘we’ per million words. In any case, even excluding Slovenia, the general frequency of ‘we’ in *Transtat*

is five times higher than in *Natstat* and *Eustat*. It is worth investigating different uses of 'we' proposed in the subcorpora.

In example (79), for instance, 'we ' is used with reference to the writer or rather to the statistical office which settles methodologies for statistical surveys as a whole. Example (79) is at the beginning of a chapter providing information on the methodology proposed as a choice made by the writer. Hence, in this case from *Natstat* 'we' is used with an exclusive function (Biber 1999: 329) which separates the writer from the readers. In examples (80) to (84) from *Eustat*, on the contrary, 'we' is inclusive since it refers to people in general and as such it refers to all human beings, including the writer and readers. This feature concurs to make the writing more personal.

The use of 'we' varies in *Transtat*, where this pronoun is used with a lot of different values:

(85) **We** are growing older. The Danish population is getting older […] (*Transtat*)

(86) **We** can forecast that the number of students will continue to decrease. (*Transtat*)

(87) **We** can find remarkable differences behind the average of the EU member states […] (*Transtat*)

(88) **We** are the first not only among EU member states, but in the whole Europe. (*Transtat*)

(89) Biodiversity includes all species on earth. Currently **we** know of about two million plant and animal species, of which 40 thousand have their habitat in the Netherlands. (*Transtat*)

(90) If **we** analyse annual changes in the number of defendants […] (*Transtat*)

(91) **We** give approximate figures. (*Transtat*)

(92) **We** classify by the amount of gross earnings all persons in paid employment […] *(Transtat)*

In example (85) 'we' can either refer to people in general or to Danish people in particular. In the former case, 'we' can be labeled as inclusive, in the latter it can be both inclusive and exclusive, because it includes all Danish people together with the writer and excludes all those who are not Danish. Since the text examined is the English translation of a Danish text, this means that the text is here addressing an international community, and not Danish people only; when the Danish yearbook was drafted in the Danish language the natural addressees were Danish people and hence a circumscribed community, which could recognize themselves as included in that 'we' which is the same of the writer . The same can also be noticed in example (88) were 'we' includes all Hungarian readers, but excludes international ones. This feature is quite interesting and involves issues related to English translations addressed to an international audience and not to a specific national community.

In example (89), 'we' is used with reference to people in general, which is comparable to the use of 'we' in *Eustat* as reported in the example.

In all the other examples, i.e. (86) to (92), 'we' is exclusive and refers to the writer, even though with slightly different values (Wales 1996). The choice of 'we' instead of 'I' by a single writer is a way to make the writing more impersonal, 'we' recalls the authority of a scholars' community and not of one single writer, in addition to that 'we' is "[…] indicative of the effort to convince the reader by emphasizing the argumentative structure of discourse" (Gotti 2006: 78). On the contrary,

inclusive ‘we’ makes the text more personal, and minimizes the distance between writer and reader, being both in the ‘we’ group.

In this specific aspect we can say that translators prefer a simpler language, overcoming the difficulty of English impersonal construction. It should also be noticed that English Translations of National Statistical Yearbooks do not change their structures and features to address the international public, but keep the characteristics of national publications. The change is in the use of English as an international language, which at times is not enough to become international.

3.3.3 Type/token ratio

The type-token ratio (TTR) is a parameter that provides us with information on the discourse of statistics. This measure studies the range of vocabulary that is used in a corpus, i.e. whether a text uses a more or less varied vocabulary than another text in the same language. Gotti (2006: 26) claims that: “The difficulty of substituting a term with its synonym has major consequences for lexical choices made in the textualization of specialized discourse and produces a certain lexical repetition.” A narrower range means the use of less varied vocabulary, which is the case of specialised texts. Biber (1999: 53) affirms that “In longer texts, there is a greater chance that words which have already been used will be repeated”. TTR is in fact very much biased by the corpus size: “TTR varies with the length of the text: longer texts have many more repeated words and therefore a much lower TTR” (Biber 1999: 53). Bias could be particularly strong for the case of the discourse of statistics which is very repetitive, and therefore the number of tokens do not increase proportionally with the number of types. For the reasons

exposed above the type/token ratio has to be calculated on similar-size corpora. In order to do that, files composing the three subcorpora under study have been randomly grouped to compare the TTR of same-size sub-subcorpora. The formula used to obtain the type/token ratio is TTR = (types/tokens) x 100:

TYPE/TOKEN RATIO			
	Types	Tokens	TTR
1)ENSY	**18806**	**1142948**	**1.6**
2) EU08	6257	77825	**8.0**
3) EU09	6610	92168	**7.2**
4) EU10	7880	122464	**6.4**
5) CY08/DK08/PT08	2466	81897	**3.0**
6) DK07/ 08/ 09	3972	102778	**3.9**
7) EST08/ 09/ 10	5170	118949	**4.3**
8) HG06/ 07/ 08	4656	95159	**4.9**
9) SK05/ 06/ 07	4208	86855	**4.8**
10) NATSTAT	3667	80259	**4.6**

Table 7 - TTR - Type/token ratio

Table 7 shows some sub-subcorpora which were created in order to have similar number of tokens. The reference number was that of *Natstat*, which is the smallest. We can see that in the corpus as a whole (Line 1), which consists of more than 1 million tokens, TTR is 1.6 and the more the token number decreases the more TTR increases. Nevertheless, some differences in the rate can be clearly detected. *Natstat* (line 10), which represents the whole subcorpora made of Ireland yearbooks, has a TTR of 4.6 which is higher than the subgroup of *Transtat* (Line 5) with a similar number of tokens and a TTR of 3. The three sub-subcorpora from

Eustat (lines from 2 to 4) have indeed the highest TTR ranging from 6.4 to 8, with the mean TTR 7. The mean TTR for *Transtat* is 4.1. Our result is that *Eustat* has the highest TTR while *Natstat* and *Transtat* subcorpora have a similar TTR ranging from 3 to 4.9. When considering TTR mean, *Eustat* keeps its position, but Natstat and *Transtat* become very similar with 4.6 and 4.1 respectively. For the purposes of the present research, we can say that the more relevant difference is between *Natstat* and *Eustat*, and between *Transtat* and *Eustat*. In our case, higher/lower TTR cannot be considered a peculiar feature of translated texts, but of specialized discourse and especially in the case of *Natstat* which, as in other cases, is the most complying with specialized discourse features. *Natstat*, in fact, uses the less varied range of vocabulary among the three subcorpora.

3.3.4 Average Sentence Length

Average sentence length is claimed to further complicate the comprehension of specialized discourse (Gotti 2006). Specialized texts use longer sentences. The assumption is that the lower the average sentence length the simpler the text. Some scholars (Baker 2001; Laviosa 1997; Oholan 2004) consider sentence length as a parameter of simplification in translated texts. Gotti (2006: 65) claims that "Written texts are encoded by far longer sentences than those found in general language". Barber (1985), who has analysed written scientific texts, calculates an average sentence length of 27.7 words. As to the present corpus, the results are similar to Barber's findings with some differences among subcorpora: *Eustat* has the longest average sentence length with 29.4 tokens and 11,072 sentences; *Transtat* has an average sentence

length of 22.1 tokens and 32,418 sentences, and *Natstat* has 20.5 as the average length and 3,846 sentences. Similarly, as to TTR, the numbers for *Natstat* and *Transtat* show slight differences also in the average sentence length, while *Eustat* has much longer sentences.

As regards the features examined, it can be said that *Natstat* adopts the most typical features of specialized discourse, while *Eustat* is the one that most differs from it. So far various differences among the subcorpora have been detected and in order to collect data for a deeper analysis of the discourse of statistics, in the following section a view into the lexis of statistical yearbooks is presented.

3.4 The lexis of statistics

The subcorpora have resulted to be characterized by a different use of some nouns and verbs which belong to the semantic field of statistics and have similar meanings. The choice for one or the other also identifies the kind of language preferred by each subcorpus.

3.4.1 ‘Number/s’ and ‘figure/s’

‘Number(s)’ and ‘figure(s)’ are two domain-specific words with high occurrences in the corpus:

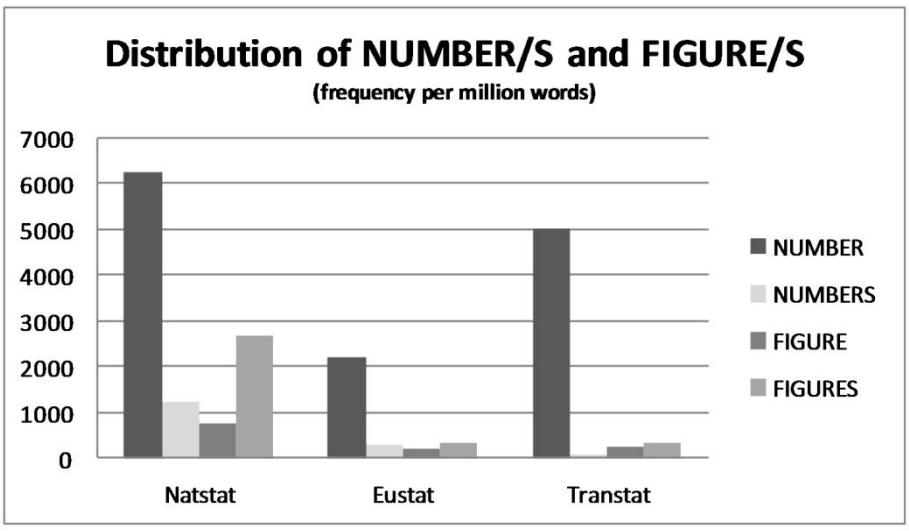

Figure 12 – Distribution of 'number/s' and 'figure/s' in the subcorpora.

In the *Transtat* keywordlist, 'number' ranks 21st when compared to *Natstat* + *Eustat*, and 11th, when compared to *Eustat*; 'numbers', 'figure' and 'figures' are not listed in the *Transtat* keyworlist in the top 100 words in both types of cross-checks.

In the keywordlist of *Natstat* with reference to *Transtat*, the word 'numbers' ranks 17th and 'figures' 59th; in *Natstat* compared to *Transtat* + *Eustat*, 'numbers' gets to 13th and 'figures' to 7th.

In the *Natstat* keywordlist, in all cross-checks a different ranking of 'number' and 'figure' can be noticed. These words are used in the plural form as synonyms of 'data', hence in some (few) cases they justify the minor use of the word 'data'. It should, however, be noticed that 'figure(s)' has a very low frequency in *Eustat* and *Transtat* compared to *Natstat*. The preference for 'number/s' can be interpreted in terms of colloquialization (Hundt / Mair 1999) thus proposing the more common and less specific word 'number'. In *Transtat* this can also be interpreted as an interference from the source languages (Toury 1995), because the use of 'number' and 'figure' is not completely overlapping with the

English one in all languages. For instance, in the case of the Italian yearbooks included in *Transtat* there are no occurrences of 'figure/s'. The word '*cifra/e*', is not always translatable with 'figure(s)' and *viceversa*. As evidence of this some examples will be reported from the Italian Statistical Abstract 2010 (*Compendio statistico italiano 2010* – Italian Statistical Abstract 2010) [42], which is not included in *Transtat* but enables comparison since it is presented online as a parallel text. The following examples show the Italian extract followed by its English translation:

(93) Qualora **la cifra** originaria sia espressa in lire […] / If **the value** is expressed in liras […]

(94) **I dati** relativi ai periodi più recenti sono in parte provvisori. / Some of the latest **figures** are provisiona**l.**

(95) **Numeri** relativi / Relative **figures**

Examples (93) to (95) confirm that 'cifra/e' and ' figure/s', even though with the same meaning do not have the same usage. This is also evidenced by the number of occurrences in the above-mentioned text (Italian Statistical Abstract 2010) where 'cifra/e' occurs 5 times and 'figure/s' 18 times. This is not the case of 'numero' and 'number' which can always translate each other, even though in some cases English native speakers would prefer 'figure' as in example (95). We can conclude that 'figure/s' is hardly ever used in *Eustat* and *Transtat* and this marks a difference in the lexis adopted by *Natstat* on one side and

[42] http://en.istat.it/dati/catalogo/20110617_00/compendio_statistico_italiano_2010.pdf (Last accessed 15 May 2012)

Eustat/*Transtat* on the other, the latter being considered to be influenced by ELF.

Also some specific words which express increment or decrement of numbers characterize the discourse of statistics. These will be analysed in the next section.

3.4.2 Increment and decrement words

Some words and connected verbs, which are the most frequently used to compare data in time and space, were studied in order to further explore the language used in the discourse of statistics. They were found to have different frequency in the three subcorpora.

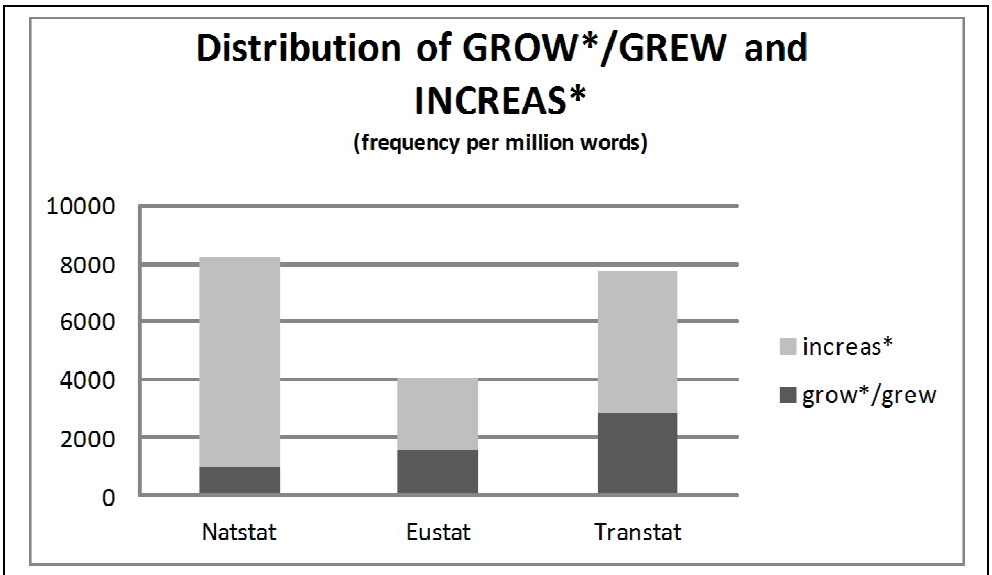

Figure 13- Distribution of increment and decrement words in the three subcorpora.

Figure 13 shows words related to 'increment' and their distribution. The Figure provides data on the use of two dynamic verbs – 'increase' and 'grow' (in all inflections) – and their respective nouns – 'increase' and 'growth'– which are all used to express similar meanings.

They appeared to have a high frequency in all their forms in ENSY and for this reason they were grouped together circumscribing our interest to their semantic field. The frequency of ‘grow*’ and ‘increas*’ is very similar in *Natstat* and *Transtat*, with a slightly lower number of occurrences in *Transtat*. *Eustat* has half the occurrences of *Natstat*. This relevant difference is consistent with data presented in this chapter where *Eustat* language appears to be characterised by fewer features of specialised discourse and more of a “story”, and in line with suggestions given in the above-mentioned publication *Making Data Meaningful* (see 1.5), which invites statisticians to write stories on numbers to make data more understandable to the lay public.

It is worth noticing that *Natstat* shows a higher frequency of ‘increas*’ words which are frequently replaced with grow*’ in *Transtat*. In *Eustat* ‘*increas’ occurrences have a similar percentage to those of ‘*grow*’. In particular, the singular noun “increase” occurs four times more frequently than ‘growth’ in *Natstat*. It is worth noticing that *Natstat* prefers ‘increase’ to ‘growth’ probably because ‘increase’ is more domain-specific while ‘growth’ is borrowed from other semantic fields. ‘Growth’ can also be used to refer to human beings, plants and all living beings in general, while ‘increase’ is only referred to inanimate agents and is therefore more appropriate to statistics. This choice can also be interpreted as preference for Latinate words as more appropriate to specialized discourse than Anglo-Saxon ones (Gotti 2006). Gotti (2006: 25) claims that monoreferentiality is one of the distinctive features of specialized discourse, and in support he quotes Piesse (1987: 58): “Never change your language unless you wish to change your meaning, and always change your language if you wish to change your meaning.”

Natstat mantains this feature and always uses the same language to express the same meaning. In ENSY as a whole, 'increas*' words are more frequent, and we can conclude that 'increas*' can be considered more typical of the discourse of statistics for both reasons expressed above. Here are examples of its use in the three subcorpora:

(96) The largest population **increases** are to be recorded in France and the United Kingdom. (*Eustat*)

(97) The most notable **increases** were in education. (*Natstat*)

(98) The proportion of the population who have been hospitalized […] **increases** with age (*Transtat*)

In *Natstat*, "increases" is used as a plural noun only in one case which is reported in example (97) while in the other instances it is used as a verb. In *Eustat*, in fact, 'increases' is used as a verb 7 times only out of 90 occurrences. All the 7 occurrences are listed in the following examples, the file reference is included to highlight repetitions:

(99) Obesity […] **increases** significantly the risk of chronic diseases. (*Eustat*_2008)

(100) Obesity […] **increases** significantly the risk of chronic diseases. (*Eustat*_ 2009)

(101) Obesity […] **increases** the risk of death and disability. (*Eustat*_2009)

(102) Life expectancy **increases** as people age. (*Eustat*_2010)

(103) Obesity **increases** the risk of chronic diseases. (*Eustat* 2010)

(104) Obesity […] **increases** the risk of death and disability. *(Eustat*_2010)

(105) The cost of teaching **increases** significantly as a child moves through the education system. (*Eustat*_2010)

The examples above list all cases of "increases" used in the singular third person simple present in *Eustat.* It should be noticed that (99) and (100) contain the same patterns, as (101), (102) and (103). Repetition of patterns is due to a typical feature of Statistical Yearbooks where the same stretches of text are often pasted year after year when updating numbers and adding new information. This feature is visible in all subcorpora and characterises the discourse of statistics analysed in this research (see 2.3).

In *Transtat*, the use of 'increases' is similar to its use in *Eustat*, hence the word mainly occurs in its plural form. Only very rarely is 'increases' a verb form, and as such it only occurs in the Denmark and the Estonia Statistical Yearbooks. For this reason the use of 'increases' as a verb in *Transtat* cannot be considered a general feature.

The prevalent occurrence of 'increase' as a noun is interpreted as a feature of nominalization which characterises specialised discourse: "This involves the use of a noun instead of a verb to convey concepts relating to actions or processes" (Gotti 2004: 58). Nominalization is also a typical feature of ELF and translated language, which is confirmed in *Eustat* and *Transtat* (see 1.6).

Also words expressing opposite meanings, namely 'decl*', 'drop*', 'decreas*' and 'fall*', were studied with regard to their distribution in the subcorpora. They resulted to be less frequent than the previous ones. 'Fall*' is homogeneously distributed in the whole corpus, while 'decreas*' has the highest frequency in *Transtat* (see Figure 14). In particular, the past participle "decreased" ranks 8 in the keywordlist

comparing *Transtat* to *Natstat* + *Eustat*, but this feature is influenced by a very high frequency of ‘decreased’ in Estonian files (70% of all occurrences in *Transtat*). Here follow examples of ‘decrement’ words and their use in the discourse of statistics:

(106) Apart from the **decline** experienced in the late 1980s, the direction of population change has since been positive. (*Natstat*)

(107) The **drop** in recipients in 1997 is a result of […]. (*Natstat*)

(108) The slight **fall** in Estonia may be due to methodological reasons. (*Eustat*)

(109) The northen countries […] experience a **decrease** in participation rates […] (*Eustat*)

(110) A slower **decrease** in net earnings was determined by a reduced income tax. (*Transtat*)

Also in the case of decrement words like ‘decrease(s)’, ‘decline(s)’, ‘fall’ and ‘drop’ the prevalence of nouns, as in examples (106) to (110), can be interpreted as a preference for nominalization. The next figure provides visual information on the use of decrement words:

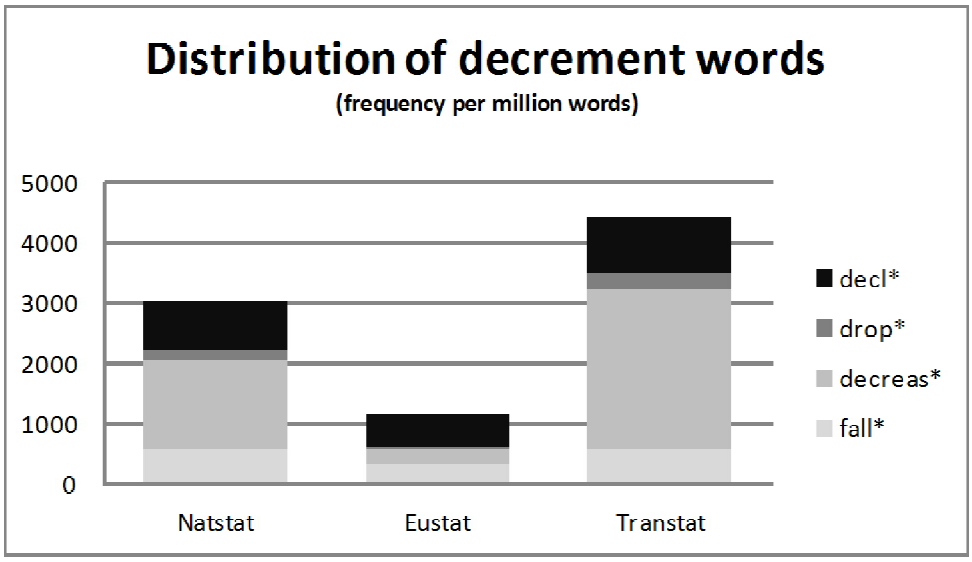

Figure 14 – Distribution of decrement words in the three subcorpora.

As can be noticed, the word 'drop*' is hardly present in *Eustat*, while it is used in *Transtat* and *Natstat.* The figure confirms the preference for 'decreas*' in *Natstat* and *Transtat* differently from *Eustat.*

3.5 Clusters and lexical bundles

Statistical language is also characterised by some clusters, which include typical words of statistical domain such as "data", "growth", "number" and "year". These clusters occur in all subcorpora though with a different frequency:

	ENSY corpus	Natstat	Eustat	Transtat
THE DATA OF	118	0	0	118
THE DATA PRESENTED	90	0	13	77
THE GROWTH OF	136	4	15	117
INCREASE IN THE NUMBER OF	124	11	5	108
DECREASE IN THE NUMBER	52	0	0	52
IN THE NUMBER OF	372	28	17	327
AVERAGE NUMBER OF	121	12	17	92

THE NUMBER OF PERSONS	166	20	16	130
NUMBER OF PERSONS EMPLOYED	73	1	16	56
DECREASE IN THE	167	11	2	154
THE PREVIOUS YEAR	534	2	6	526

Table 8 – ENSY most common clusters, and their occurrences in the three subcorpora.

It should be noticed that some clusters do not occur in all subcorpora; that is why they are excluded from the following figures, which only show the frequency of the above-mentioned clusters per million tokens:

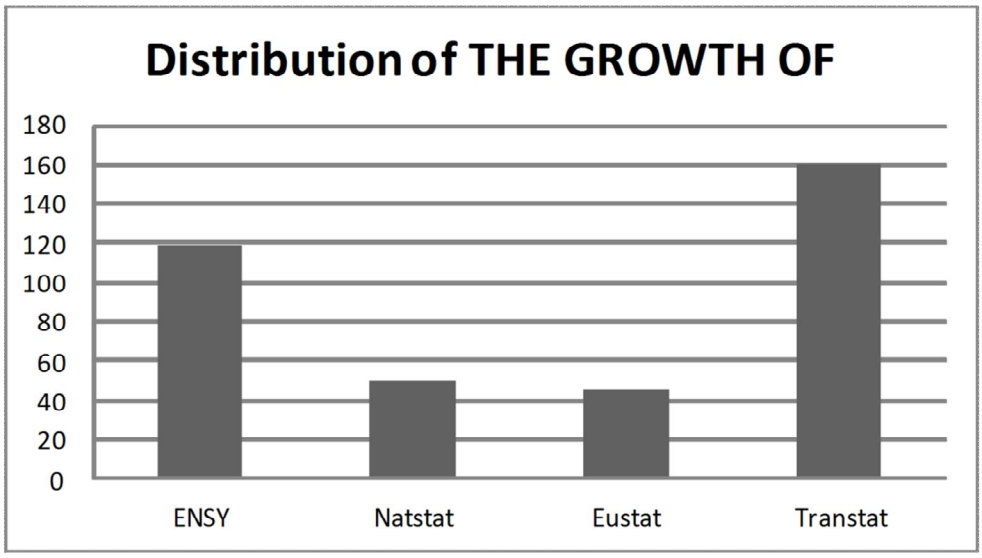

Figure 15 – Distribution of 'The growth of' in the three sucorpora.

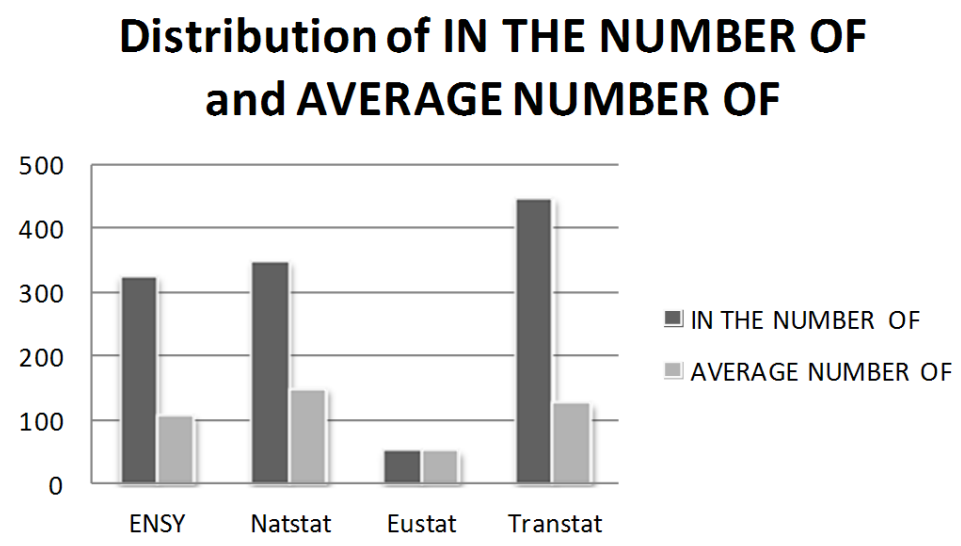

Figure 16 – Distribution of 'in the number of' and 'average number of' in the three subcorpora.

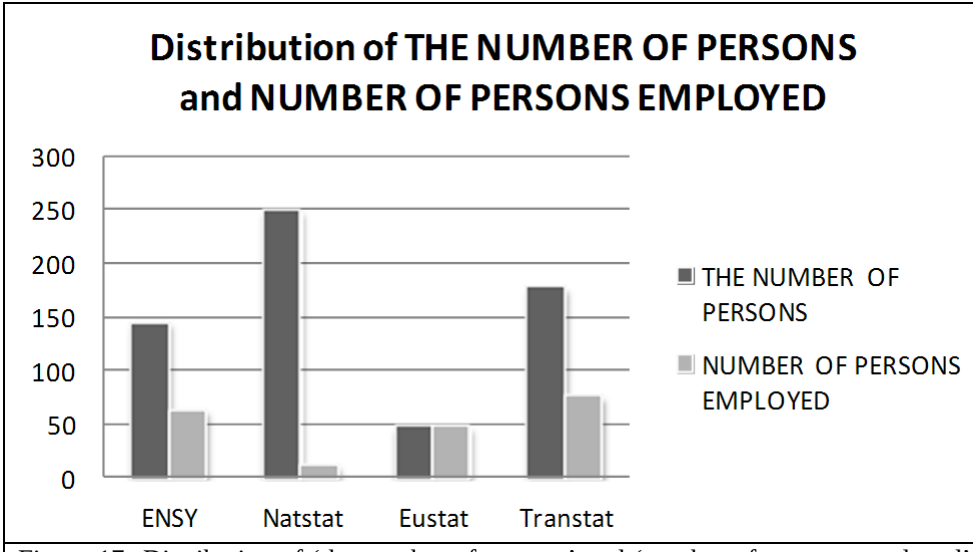

Figure 17- Distribution of 'the number of persons' and 'number of persons employed' in the three subcorpora.

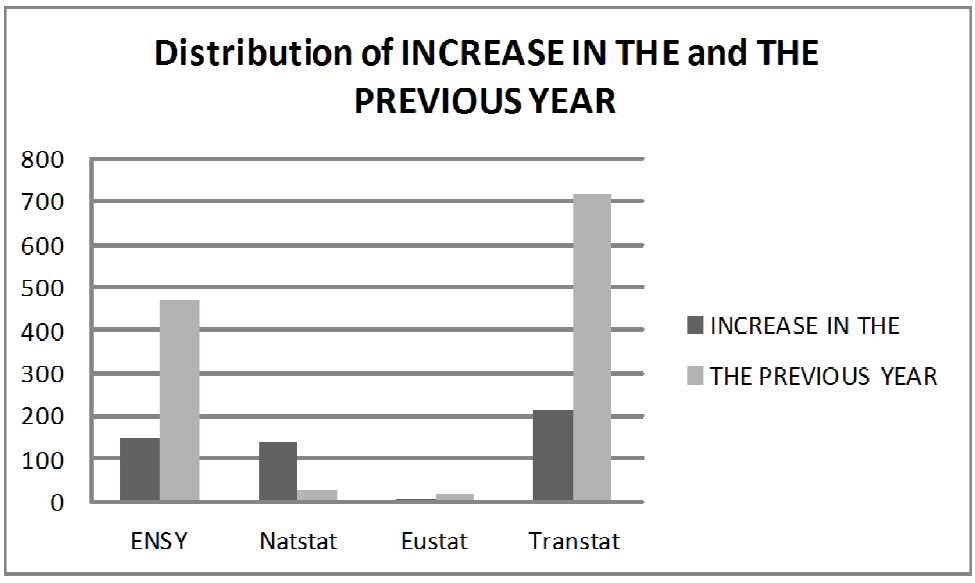

Figure 18 – Distribution of 'increase in the' and 'the previous year' in the three subcorpora.

Generally speaking, three-word and four-word clusters are quite frequent in the discourse of statistics, the highest frequency being recorded in *Transtat* and the least in *Eustat*.

The most frequent clusters in each subcorpus, containing 'year' from minimum 3 to 5 words, are the following:

(a) in recent **years** (55 occurrences in *Eustat*)

(b) aged 15 **years** and over (20 occurrences in *Natstat*)

(c) the previous **year** (526 occurrences in *Transtat*)

It is worth noticing that the 4 most frequent 4-word clusters of 'year' in *Transtat* always include 'previous' (see 3.2.3.1).

The observation of these clusters led us to look at them as possible lexical bundles in the discourse of statistics. In LGSWE, Biber (1999:

990) defines a lexical bundle “a recurring sequence of three or more words”. As he remarks, “Lexical bundles can be regarded as extended collocations: bundles of words that show a statistical tendency to co-occur” (Biber 1999: 990). Biber identifies a certain number of occurrences per million words as a parameter for recognising lexical bundles: three-word sequences are set a minimal cut-off of at least twenty times per million words, and four-word sequences a cut-off of at least ten times per million words. As for our corpus (1,142,948 tokens), all the mentioned clusters can be said to have more than enough occurrences and could be regarded as lexical bundles of statistical texts.

Biber (1999: 992) notices that “shorter bundles are often incorporated into more than one longer lexical bundle”. This is also the case of statistical texts:

‘the number of’ is part of ‘increase in the number of’,’ in the number of’, ‘the number of persons’;

‘number of persons’ is part of ‘number of persons employed’.

Another interesting feature analysed by Biber (1999: 991) is that “In academic prose lexical bundles are more commonly parts of noun phrases and prepositional phrases”. Also in the present study, we found incomplete lexical units which are part of noun phrases, such as ‘the number of persons’, ‘increase in the number of’.

In LGSWE, some clusters are identified as lexical bundles of academic prose and are presented as examples:

(d) ‘there was no significant’

(e) ‘in the case of the’

(f) ‘it should be noted that’.

These lexical bundles were searched in ENSY with the following results:

(d) – 5 occurrences; (e) – 4 occurrences; (f) – 35 occurrences.

Hence, these lexical bundles can also be considered recurrent in ENSY. This means that this specific feature of academic prose also occurs in the discourse of statistics. The more numerous occurrences (24) of (f) are characteristic of *Eustat.* The language of *Eurostat* tries to guide the reader through the interpretation of data, and the lexical bundle ‘it should be noted that’ constitutes a typical introduction to it.

3.6. The case of ‘example’ on the way to Clarity

As already mentioned, *Eustat* is characterised by particular features in its effort to be clear and reader-friendly; this section aims to analyse this aspect more deeply by investigating the word ‘example’.‘Example’ is the 18th in the keywordlist of *Eustat* when compared to *Natstat*+*Transtat*, and maintains a high rank also when *Eusta*t is compared to the two subcorpora separately. It is an interesting word that tells us how some data are connected to reality by means of examples. It is quite rare in *Natstat* and *Transtat.* The most common left collocate of ‘example’ is ‘for’. The following figure shows the distribution of the cluster ‘for example’ per million words:

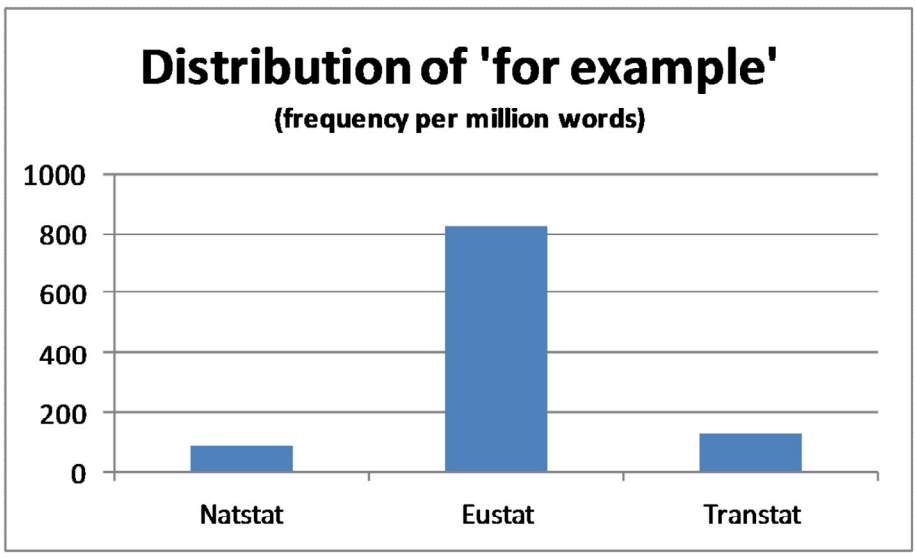

Figure 19 – Distribution of 'for example' in the three subcorpora.

Some examples of the occurrences of the cluster 'for example' can help understand its use in *Eustat*:

(111) […] other current transfers, **for example** workers remittances. (*Eustat*)

(112) […] one group of indicators relate to monetary (income) poverty analysed in various ways (**for example**, age, gender, activity, status). (*Eustat*)

(113) **For example**, reliable statistics are needed to asses macro-economic developments. (*Eustat*)

(114) […] integration of data from many sources, **for example**, statistical surveys of business and households and administrative data. (*Eustat*)

The high difference between the use of 'for example' in *Eustat* compared to other subcorpora has led us to check for the presence of other synonyms in *Natstat* and *Transtat*. All other possible references to examples where searched, namely 'for instance', 'ie', 'i.e.', 'eg', 'e.g.'.

'Ie' and 'eg' were considered both with and without dots, because they resulted to be only followed by dots in *Eustat* and *Transtat*. Not all the above-mentioned expressions have exactly the same meaning: 'e.g.' is the abbreviation of Latin *exempli gratia*, and hence should be always used to introduce an example, being the most similar to 'for example'; 'i.e.' is the abbreviation of Latin *id est*, therefore it introduces a further and more specific explanation, as it has the meaning of 'that is' and provides all-inclusive information on the issue it refers to, and not partial as is the case of examples. The cluster 'for instance', which could replace 'for example', is very rare in the corpus (0% in *Natstat*, 0.004% in *Eustat*, 0.005% in *Transtat*), hence it cannot be considered relevant for the comparative analysis.

The following Figure reports the distribution of the expressions presented above:

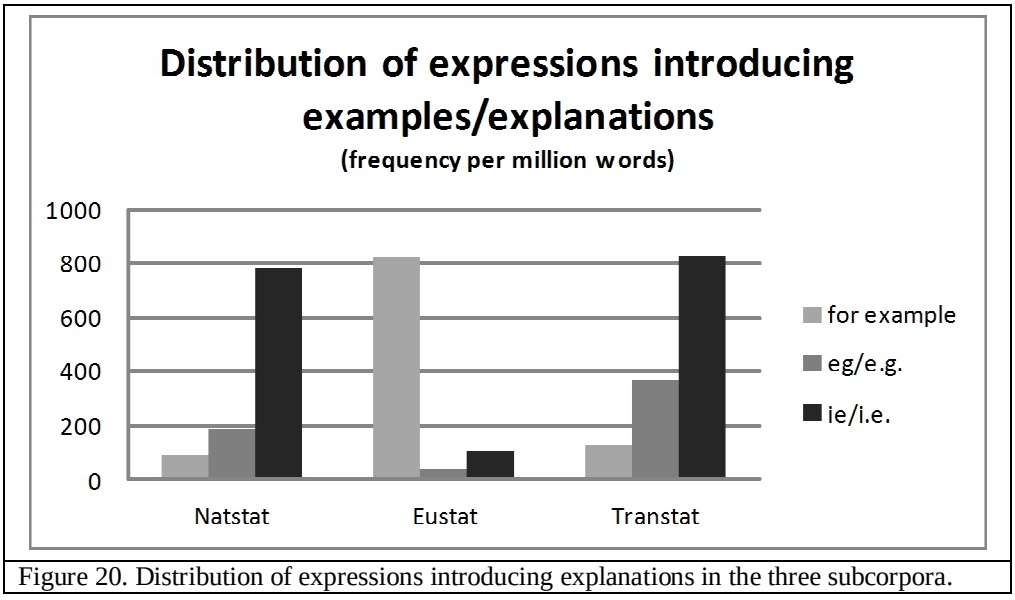

Figure 20. Distribution of expressions introducing explanations in the three subcorpora.

The differences across subcorpora are very sharp. In *Eustat*, 'for example' is ten times more frequent than in *Natstat* and seven times

more than in *Transtat*. The difference is only partially compensated by the use of 'e.g.'; *Transtat*, prefers this expression when introducing examples. In *Natstat*, 'eg' has a very low frequency, but double than 'for example'. These findings reveal that *Natstat* and *Transtat* provide a small number of examples to the readers and, when they do so, they prefer introductory 'eg'. This feature could confirm the phenomenon of "colloquialization" (Hinrichs / Szmrecsanyi 2007) already detected in *Eustat* for the use of *s*-genitive and of 'semi-modals'(see 3.1.1). As to *Transtat*, it evidences the feature of 'conservatism' (Baker 1996), which is the tendency of translated texts to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them. Here are some examples on the use of 'eg' and 'e.g.' in *Natstat* and *Transtat*:

(115) Figures exclude non-criminal prisoners (**eg** immigration detainees) and those on trial and on remand. (*Natstat*)

(116) […] usually the year previous to the benefit data **eg** 2008 figures use the 2007 mid year estimates. (*Natstat*)

(117) […] the cost of non-industrial services rendered by others (**e.g**. telephone, telegraph, telexes and postage charges, advertising, legal services, accounting and auditing, insurance etc.). (*Transtat*)

(118) […] contribution to semi-budgetary organisations (**e.g.** in education). (*Transtat*)

(119) In addition to this the Danish state pays the expenditure on operating, **e.g.** the legal system and defence. (*Transtat*)

As can be noticed in examples (115) to (119), 'eg' and 'e.g.' are always used to introduce examples of more general categories in order to make them understandable to the readers. It should also be noticed that in

Natstat ‘eg’ appears four times only in each yearbook (2007, 2008, 2009) to introduce examples; in the 2010 yearbook, instead, it is used only twice and is written ‘e.g.’. We can infer that examples are not considered relevant in the presentation of Irish statistics, which compose the whole *Natstat* subcorpus. On the contrary, ‘ie’ is much more used and has a similar frequency to ‘i.e.’ in *Transtat.* In this section we cannot consider ‘ie’ as a synonym of ‘for example’ or ‘eg’ but as one of the strategies used to make statistical texts (somehow) clearer and more accessible. Some examples of ‘ie’ and ‘i.e.’ occurrences are the following:

(120) Fixed assets acquired from others were valued at the full cost incurred **i.e.** at the delivery prices plus installation costs. (*Transtat*)

(121) In 2007, the production of fruits and berries was five kg per inhabitant, **i.e.** bigger than in the previous year. (*Transtat*)

(122) […] primary school teachers diminished by more than 5 thousand, **i.e.** nearly 7%. (*Transtat*)

(123) […] only when a transfer of ownership occurs **ie** when payment is received. (*Natstat*)

(124) […] by the main demographic characteristics, **ie** age, sex, and marital status. (*Natstat*)

(125) […] the number of very young persons (**ie** aged 0-4). (*Natstat*)

The word ‘example’, so frequent in *Eustat*, was also studied in a diachronic perspective in this subcorpus. *Eustat* was therefore subdivided into two subcorpora, the former from 2006 to 2008 and the latter from 2009 to 2010. The result, which was calculated per million words, is shown in Figure 21:

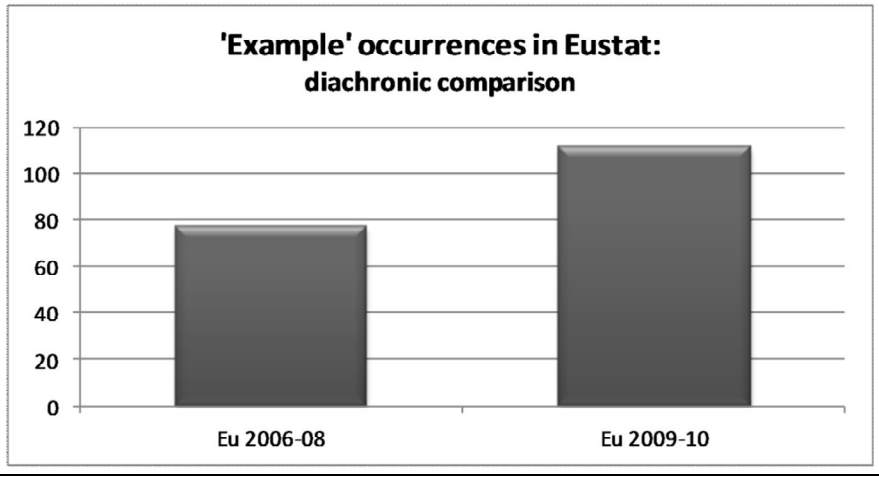

Figure 21 – Diachronic comparison on occurrence percentage of 'example' in Eustat.

It is worth noticing that *Eurostat* has increased the number of examples in the latest years and this increment is to be ascribed to its efforts for improving Clarity. Examples are a means of communication with the international public and facilitate readability by people who are not expert in statistics. This is very interesting and qualifies Eurostat on the way to clarity as is required by Regulations. It also meets the needs of readers who want to approach statistics and comprehend what they are reading.

Eustat is characterised by some grammatical phrases which make the text clearer. Murphy (2008) notices that EU editors, who revise texts drafted in English, do not delete expressions like 'in order to' and 'as well as' in edited texts. She interprets this as a way of making texts clearer. This feature is also quite common in *Eustat.* The expressions 'in order to' and 'as well as' are present in Eustat more than in the other subcorpora as shown in the following Figure:

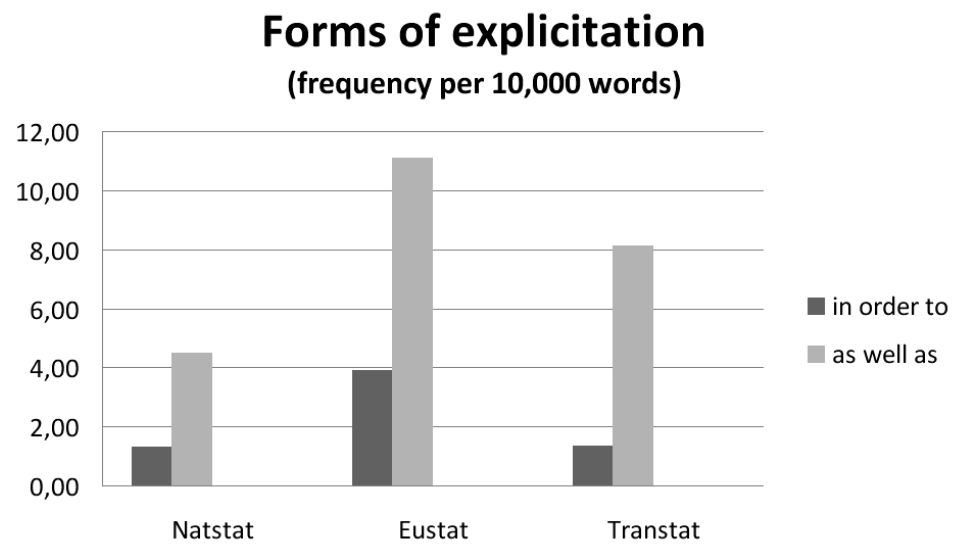

Figure 22 – Distribution of 'in order to' and 'as well as' in the three subcorpora.

From the data collected and also from the above-exposed remarks, *Eustat* results to be the ENSY subcorpus which best meets the requirements of clarity at least in some aspects. The efforts made by Eurostat to disseminate statistics at the European level could guide NSIs on the way to clarity.

In the Conclusions, the research provides some new insights into this aspect also as a contribution to the development of international communication in statistics.

4. ENGLISH AND EUROPEAN STATISTICIANS

This chapter presents the results of a survey that was carried out to test European statisticians' perception on issues related to the discourse of statistics. The aim was to analyse if and how statisticians have implemented the new Regulation on Clarity and Accessibility. The survey proposes very simple and easy-to-answer questions. The questionnaire, which is explained in detail in section 4.2, is visible in Annex 1. As underlined in the previous chapters, (see 1.6) care for language is a new entry in the scientific community of statisticians, and awareness of the topic is circumscribed to some European countries. Data on the respondents' proficiency in English, their educational background, their familiarity with statistical publications in English were collected to draw a picture of the context of statistics discourse.

4.1 Questionnaire design

The questionnaire is divided into two sections. The first section contains data on respondents, including the field of study, educational level, English proficiency, the field of statistics in which they work and the occasions when they were required to use written or spoken English to communicate with one another (i.e. conferences, abstracts, working groups, etc.). Such data are functional to drawing comparisons among different groups of specialized statisticians (i.e. the economist, social methodologist statisticians etc.). The second section focuses on Statistical Yearbooks in English: their readability, clarity, and feedback from users. Respondents were also required to answer multiple-choice questions and fill-in blank spaces to propose suggestions for improving clarity.

When some answers are missed by respondents we can speak of 'partial non-response'[43], although we have a very limited rate of such type of response and only in two questionnaires.

4.2 The sample

The questionnaire was sent to all European National Statistical Institutes (NSIs) of the EU member states, Eurostat and an Italian University. Thus, the total number of collected and filled-in questionnaires is 43: the National Statistical Institutes of Italy, Bulgaria, Czech Republic, Cyprus, Estonia, Finland, Ireland, Lithuania, Romania, Slovenia, Hungary, Sweden and Switzerland (the last as member of EFTA – European Free Trade Association), Eurostat, and the University of Roma "La Sapienza". Not all respondents are statisticians, or better not all of them have a university degree in statistics, but all work in statistics fields. In some cases, such as Eurostat, Lithuania, Czech Republic and Slovenia, respondents work in departments for statistics dissemination; in others they are in charge of sections of international departments and therefore they are well aware of the issues proposed by the survey, as confirmed by findings.

The majority of respondents have a high educational level, as results from 18 university degrees, 18 master's degrees and 7 PhDs; this means that about 60% of the interviewed statisticians have achieved higher qualifications. Their field of study is mainly scientific (i.e. Statistics, Economics and Mathematics); those who studied humanities account for 27% only. This is exactly the same percentage of respondents who work in Communication and international relations.

[43]For the definition see: http://stats.oecd.org/glossary/detail.asp?ID=3764.(Last accessed May 2012)

As underlined in section 2.7.1, not all European NSIs have replied by filling-in the questionnaire. It is worth noticing that not all European NSIs publish the Statistical Yearbook in English (see 2.2); furthermore, some NSIs have a very small number of publications and data published and translated into English (i.e. France, Germany, Austria, Belgium, Luxemburg, Spain). The countries that publish the English translation of their Statistical Yearbook, either on-line or on paper, are 16 (i.e. Hungary, Slovenia, Slovakia, Czech Republic, Portugal, Poland, The Netherlands, Lithuania, Latvia, Italy, Greece, Finland, Estonia, Denmark, Cyprus and Switzerland); of these only five did not send the questionnaire back. The findings reveal that the large majority of NSIs who have an international policy to foster accessibility and reach out an international public have participated in the survey.

The data collected by means of the questionnaire refer to June 2011. As to respondents, they are 29 females and 14 males, in their majority (84%) aged under 50, hence the sample includes young and senior statisticians. The distribution of respondents is reported in Figure 23:

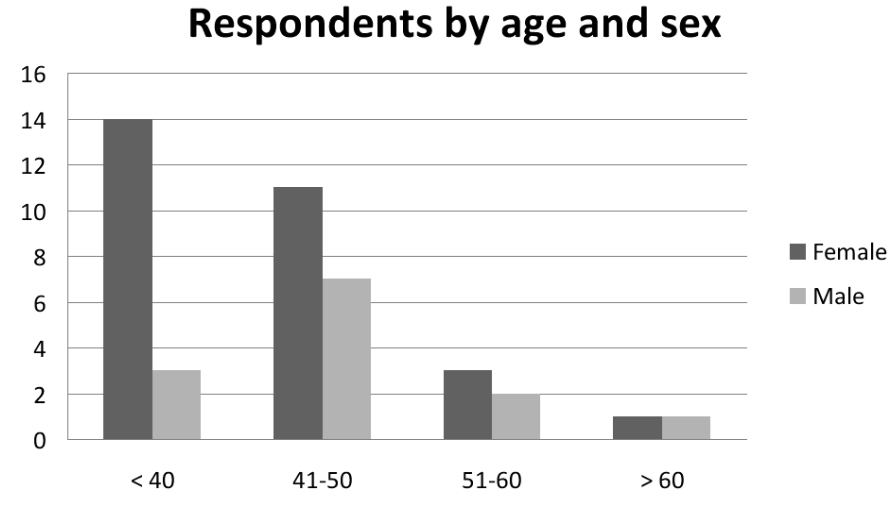

Figure 23- Distribution of respondents by age and sex.

All the statisticians interviewed work in the same country in which they were born, hence they usually use their native language at work. Exceptions are: a Spaniard working in the Ireland NSI, a Russian woman working in the Lithuania NSI (although this is due to historical reasons, since Lithuania was part of the former Soviet Union), and the three respondents working at Eurostat (i.e. Belgium, Italy and Finland). The only respondent who is an English native speaker is the Irish one. Therefore, we can say that our sample is composed of non-native English speakers, whose majority (72%) declare to possess an advanced or excellent level of proficiency in English. One more common feature is that all respondents (100%) answered 'yes' to the question "Have you ever used English to communicate on statistical topics in the last five years?". This confirms the relevance of the English language in the scientific community of European statisticians. They all use English whatever their task is within the institution and 90% of them use English both in speech and writing, at least on two of the eight specified as possible answers (26 abstracts; 29 papers; 21 articles; 22 publications; 18 readings; 30 meetings; 27 conferences; 9 classes). As a result, the use of English appears to be connected to official occasions and specific statistics issues.

4.3 Findings

This section presents the data collected to test the respondents awareness of Accessibility and Clarity by means of a questionnaire. The interpretation of the findings, however, will be discussed in detail in the

following section. In the meanwhile, the presentation of questionnaire responses will be provided by grouping them homogeneously.

All respondents affirm that “Accessibility and Clarity” also refer to the language used to present/explain statistics data, even though in informal conversations they admitted they did not know about the specific Regulation on the matter.

Only three of them consider “Language of texts more relevant than table presentation”, nine admit that “Language of texts is more relevant than table presentation”, and the majority (72%) attributes equal importance to tables and language in texts. Those who chose the option “Table presentation is more relevant than language texts” have all an economic or statistical educational background, and are probably more accustomed to the non-verbal elements (cf. Widdowson 1979).

Again, 72% of the respondents (31 out of 43) consider that the new requirements of Accessibility and Clarity to reach out non-expert users and the wider and international community are leading to language simplification. Only ten of the statisticians interviewed responded that this was not the case, and they have in common the field of statistics in which they work, i.e. social statistics.

In question 4 of Section 2, statisticians were asked to rank with respect to quality and within a range from a minimum of 1 to a maximum of 10 the relevance of the language used to present/explain statistical data. As a result, language was considered very important for achieving quality, since 23 ranked language 8 and over, as represented in Figure 24:

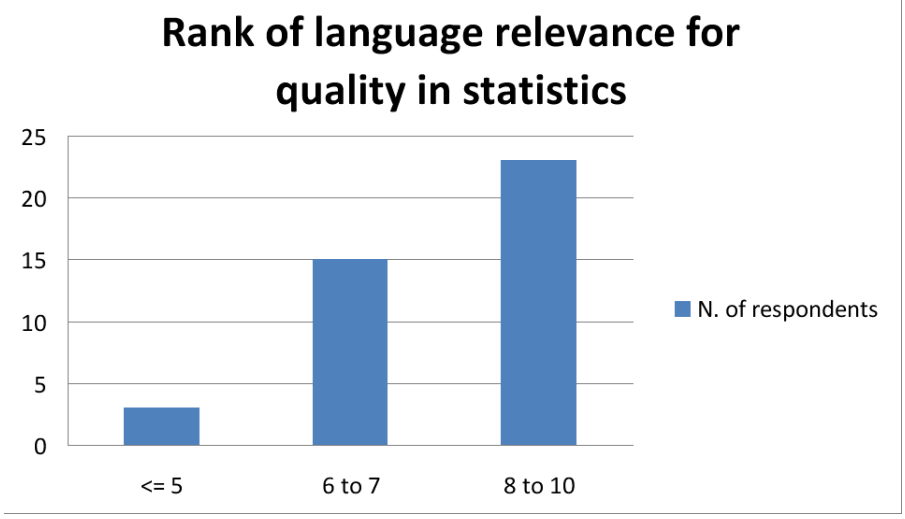

Figure 24- Relevance of language for quality in statistics.

What is interesting here is that even though Clarity and Accessibility, as well as access to the international community, are given prominence by the Code of Practice (2005) and Regulation 2009, very few among the interviewed statisticians formally request users to send a feedback on the topic. Only 12 out of 43 admitted they have a feedback by users on the clarity and accessibility of publications translated into English. Even some respondents working with data dissemination admitted they had no feedback at all by users whether these found statistics texts clear and accessible or not.

The last part of the questionnaire focuses on statistical yearbooks. The majority (68%) of respondents appear to have read Eurostat Statistical Yearbook in its original English version. About one third of respondents (36%) find Eurostat Statistical Yearbook not clear enough. Since respondents are statistics expert-users and peer members of the same scientific community, their negative response is estimated to have a higher rate among non-expert readers.

Among the possible suggestions for improving clarity in Eurostat Statistical Yearbook, the majority of respondents chose “Provide additional explanations when referring to typical national phenomena” as their first option; “Improve glossaries” as second and “Include a greater number of examples” as third. Figure 25 provides a complete picture of the preferences expressed by respondents:

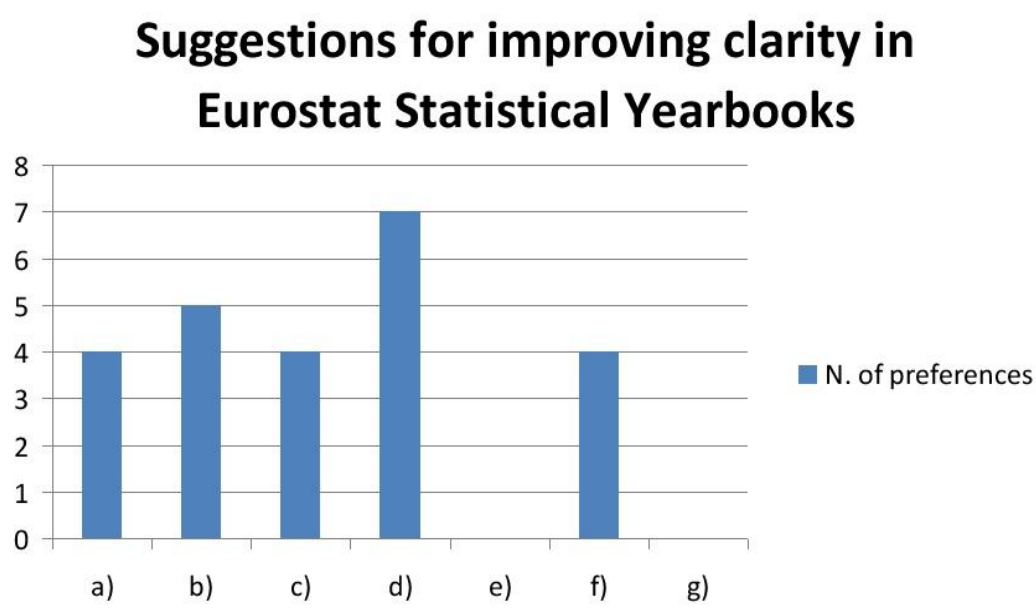

Figure 25

Legenda:

- **a)** Include a greater number of examples
- **b)** Improve glossaries
- **c)** Shorten sentences
- **d)** Provide additional explanations when referring to typical national phenomena
- **e)** Use a different lexicon
- **f)** Use more visual representations
- **g)** Other (specify).

The questionnaire also includes a question on clarity with reference to the English translation of statistical publications produced by the institutions for which respondents work. The possible answers are:

"Yes", "No", "Do not know". The findings show a high rate of "Do not know". This seems a politically correct answer not to offend the organization for which they work, and some respondents even asked for the questionnaire to be anonymous. Figure 26 below reports the statisticians' opinion on Clarity in English translations produced by their bodies. As noticed above, all answers are to be interpreted in the perspective of respondents who are expert users of statistical publications. There is evidence that even statisticians consider their publications not clear enough, and in some cases (7) not clear at all. We could infer that the use of the English language is not accompanied by competence in writing statistical official publications in English:

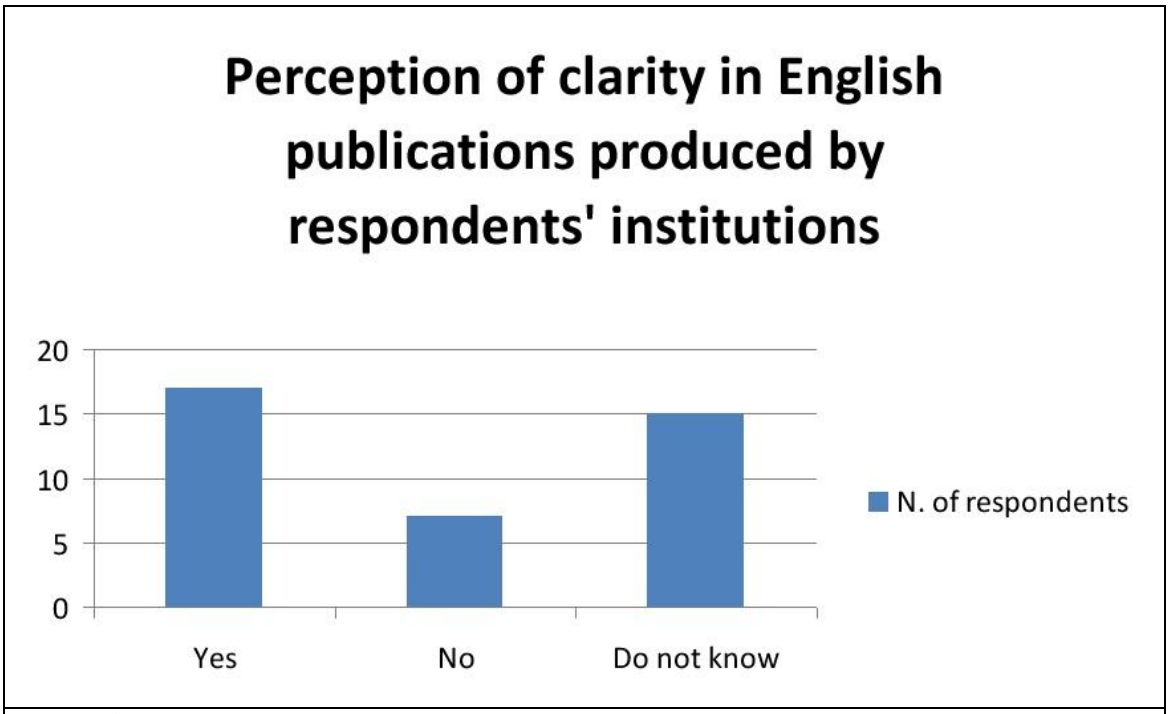

Figure 26

Figure 26 shows that the perception of clarity is quite low. It should also be noticed that respondents were referring to publications produced by the institution for which they work.

When referring to the English translation of a member-state Statistical Yearbook, only twenty respondents admitted that they had read it and half of them found it not clear enough. The fact that 23 respondents provided suggestions for improving clarity in translated texts means that a large number of respondents found translations not clear enough. Among their suggestions are: first, “Further revision of English texts by expert statistician with high English proficiency”, second, “Provide additional explanations when referring to typical national phenomena” and “improve glossaries”; third, “Produce an original English version in place of a translation” and “shorten sentences”.

In the following Figure, all suggestions proposed are reported on the grounds of preferences expressed by respondents:

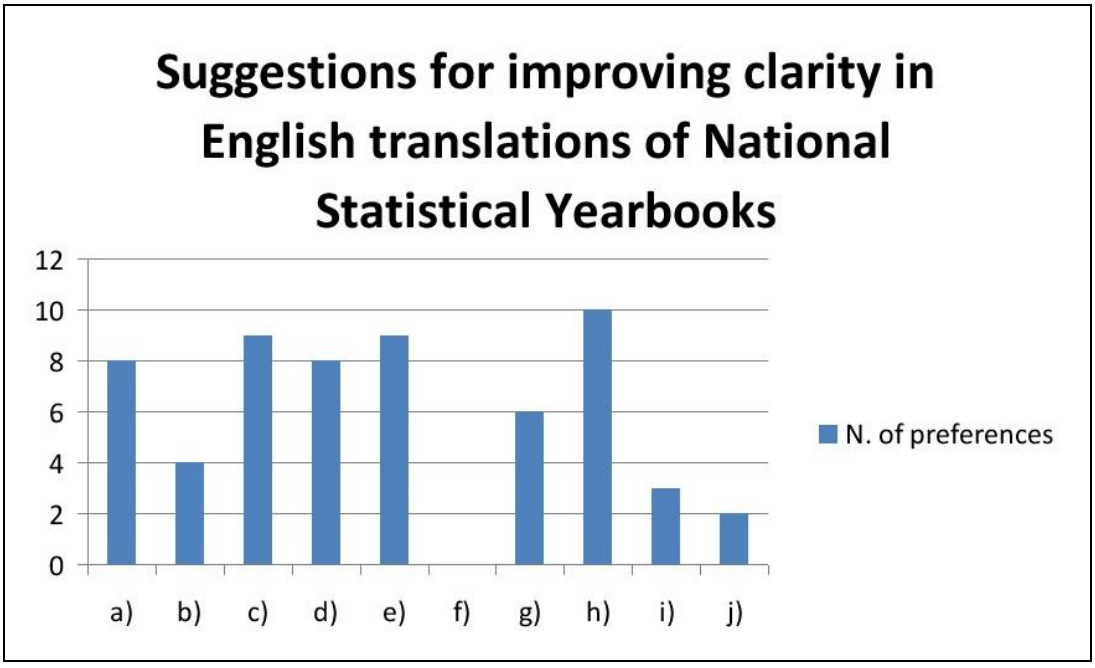

Figura 27

Legenda:

a) Produce an original English version in place of a translation
b) Include a greater number of examples
c) Improve glossaries
d) Shorten sentences
e) Provide additional explanations when referring to typical national phenomena
f) Use a different lexicon
g) Choose more expert translators
h) Further revision of English texts by Expert statisticians with high English knowledge.
i) Use more visual representation than in the original text
j) Other (specify)……..…..

The most recurrent suggestion is *h) Further revision of English texts by Expert statisticians with high English knowledge*. This means that the respondents find the use of the English language not always correct and understandable. It should however be noticed that a good number of respondents chose *a) Produce an original English version in place of a translation*. This is interesting for the point of view of this study since it underlines that a mere translation of statistical texts does not make them accessible and clear to the international public unless it is more deeply revised.

4.4 Comments on data

As mentioned above, the survey is an attempt to provide information on the perception of European statisticians of the use of the English language in statistical publications although on a circumscribed sample. The survey led statisticians to reflect on a topic that is new to them and that is not within their usual professional concerns, as they reported in informal conversations when the survey was submitted.

The data collected confirm that English is used as a Lingua Franca in official and academic statistical settings by highly educated people with different levels of proficiency in English, which is in line with what is claimed by ELF scholars on the use of English by non-native speakers (Seidlhofer (2005); Jenkins (2003); House (2004); Taviano (2010)) and its widespread use. They also confirm that English is the communication language among Europeans in the discourse community of statisticians (Swales 1990) regardless of the field of statistics in which they work.

Furthermore, all respondents consider Clarity and Accessibility also related to the language used to present statistics (see question 2.1 in the questionnaire). This is worth noticing because clarity and accessibility could also be referred to the means used for the dissemination of statistics (e.g. the Internet, paper publications, etc.), or the non-verbal elements (e.g. tables and graphs). This answer points to the relevance that statisticians begin to attribute to language and verbals accompanying the presentation of table and graphs which are considered to have an equal standard of importance (see question 2.2 in the questionnaire). This increasing interest for language has to be connected to the simplification of statistical language which respondents affirm is taking place in this specialized domain.

Another interesting aspect resulting from data is the relationship between quality and language. Quality in statistics is defined by Eurostat as resulting from the simultaneous co-working of 6 criteria (cf. 1.7) :

> 1. Relevance: an inquiry is relevant if it meets users' needs. The identification of users and their expectations is therefore necessary. In the European context, domains for which statistics are available should reflect the needs and priorities expressed by the users of the European Statistical System (completeness).
>
> 2. Accuracy: accuracy is defined as the closeness between the estimated value

> and the (unknown) true value.
>
> 3. Timeliness and punctuality in disseminating results: most users want up-to-date figures which are published frequently and on time at pre-established dates.
>
> 4. Accessibility and clarity of the information: statistical data have most value when they are easily accessible by users, are available in the forms users desire and are adequately documented.
>
> 5. Comparability: statistics for a given characteristic have the greatest usefulness when they enable reliable comparisons of values taken by the characteristic across space and time. The comparability component stresses the comparison of the same statistics between countries in order to evaluate the meaning of aggregated statistics at the European level.
>
> 6. Coherence: when originating from a single source, statistics are coherent in that elementary concepts can be combined reliably in more complex ways. When originating from different sources, and in particular from statistical surveys of different frequencies, statistics are coherent in so far as they are based on common definitions, classifications and methodological standards.[44]

In n. 4 we are reminded that "statistical data have most value when they are easily accessible by users" and in n.1 that users' needs are essential in preparing statistical data. Hence users and quality are interrelated especially in the Eurostat perspective. It is worth noticing that clarity is also focused on by Eurostat as a way to meet users' needs. These elements concur to recognise that quality and clarity in statistics are more and more interrelated and lead to issuing guidelines on clarity aimed at simplifying the language of statistics.

The lack of feedback by users is indeed limiting the possible development of more accessible publications. Publications in English, which are addressed to an international public, should undergo a deeper study to point out their limits and their degree of clarity and accessibility. In this approach the suggestions provided by respondents to this survey

[44] Eurostat, "Assessment of quality in statistics - Definition of Quality in Statistics", Working Group, Luxembourg, October 2003. http://stats.oecd.org/glossary/detail.asp?ID=2215. (Last accessed 3 June 2012)

could be a good starting point for a fresh way of understanding statistical publications in English.

When referring to Eurostat Statistical Yearbook no remark on the language was found. None required a proof reading of Eurostat publications. Respondents claimed for some additional explanations when referring to national phenomena and for improved glossaries. These aspects are indeed relevant for users in order to facilitate their understanding. Different national aspects need additional explanations so as to enable readers to enter a different culture and view of the world, lifestyle and national, social and economic contexts. The request for more examples, which is the third with reference to the Eurostat Statistical Yearbook, has been partially implemented (see question 3.6 in the questionnaire) as a strategy used by Eurostat to make statistics more accessible to a large public.

In question 2.13, suggestions are made with reference to national yearbooks translated into English; in this case, a simplified English language is presented as a necessity. This is also remarked in the blank space in which "more English translation" is required, which confirms what Taviano (2010: XIV) claims: "[…] an analysis of the complex relations between ELF, translation and globalization clearly testifies the fact that, rather than disappearing, translation and translators have a strategic role, perhaps more than ever before". The blank space left for other possible suggestions was filled in, among others with "proofreading by a native English speaker", in addition to respondents who chose *h) Further revision of English texts by expert statisticians with high English knowledge*. Such remarks emphasize that European National Statistical Yearbooks translated into English need further

editing to make the English language used more accessible and clear.

Even for national yearbooks improvement by means of glossaries is considered vital. It is particularly worth noticing that many suggest "produce an original version in place of a translation". This suggestion should be taken into account since the English translation of National Statistical Yearbooks may be difficult to be interpreted by foreign readers. If users' needs should shape data presentation, some reflections on interrelationships between content and language should be made. Since each National statistical Yearbook is conceived of as a text to be read by the national public in the national language, when it is translated into English some mediation is required. Different readers require further explanations, or simplifications. Something is done by translators who are always mediators (Baker 2011)[45], still much has to be done. Whenever a text is written, readers' needs should be taken into account and feedback is necessary to facilitate efficient communication; the very limited cases of feedback reported by the surveyed statisticians are evidence of the limited policy implemented in the field of improving clarity at European NSIs level.

[45] http://manchester.academia.edu/MonaBaker/Papers/149075/Ethics_of_Renarration. (Last accessed 26 May 2012)

5. CONCLUSIONS

This study has investigated a specific type of English language used by Eurostat and by EU NSIs to draft Statistical Yearbooks.

The research has pointed out how disseminating statistics in an efficiently communicative way is topical at this specific time and in this world where National borders are permanently overcome.

The guidelines provided by European bodies to make statistics clear and accessible are calling NSIs to communication strategies which should enable specialized and non-specialized readers to read and understand European national statistics.

Answers to the research questions (see 2.1) are provided by an interpretation of the findings collected from an analysis of the ENSY corpus and its three subcorpora. An interpretation of the findings reveals the main features of the three subcorpora as follows.

The *Transtat* subcorpus is characterized by: anaphoric reference, evidenced by the high frequency of 'the' referred to an already specified element; preference for *that/those+ of* phrases to the elliptic genitive, which is used to provide an explicit reference; preference for 'of' instead of *s-genitive,* which is extremely rare (it was found in *Eustat* only); a high frequency of 'of' and 'in', opposed to the use of nouns and participles with an adjectival function in the other subcorpora. This feature is particularly evident in reference to the nouns 'year' and 'data'. The prepositions mentioned above are used to specify the type of data and the year of reference.

The *Transtat* subcorpus is also characterized by explicitation, disambiguation, and conservative language among the Universals of Translation.

Furthermore, the use of 'we' has been detected in place of passive impersonal forms, along with a very a high frequency of three to five-word clusters.

The main features of *Eustat* can be summarized as: a very high frequency of examples (examples are very rare in *Natstat* and in *Transtat)*, which make the text more accessible and clear. In *Eustat* the majority of examples are introduced by the cluster 'for example'. An increasing frequency of examples has been detected in the latest years.

The *type/token ratio* is the highest of the three subcorpora, and therefore a more varied vocabulary is used. Sentence length is the highest.

The use of 'can' is preferred to 'may'. The latter is more frequent in *Natstat* only. The preference for 'can' is one of the elements of colloquialization of Eurostat statistical discourse together with the use of the *s-genitive*.

It is worth noticing that both in *Eustat* and *Transtat* nominalization (preference for nouns instead of verbs) is more used, which is typical of ELF.

In the *Natstat* sub-corpus the following features were detected: traditional patterns of specialized discourse, such as, for example, passive and impersonal forms, or the use of Latinate words (e.g. preference for 'increase' instead of ' grow'); preference for 'may' in place of 'can', the latter being considered more colloquial and preferred in the other two subcorpora; very rare exemplifications; an extremely reiterative language.

As mentioned above, some ELF features have resulted to characterize both *Eustat* and *Transtat*, namely nominalization, colloquialization and explicitation. Therefore, some similarities can be detected between the English of translated texts and the EU language aimed at making statistical texts accessible and clear to the readers. It should be noticed that both *Eustat* and *Transtat* statistical texts are targeted at an international public. The aim of reaching out European expert and non-expert readers gives room to the adoption of language features which fall in the field of ELF. In addition, features of colloquialization, which were mainly detected in *Eustat*, were also found in some of the *Transtat* files and can be considered forms of language simplification. It could be said that even though the goals of Accessibility and Clarity are still far to be reached, the interest for the language of statistics is increasing and is giving room to a new approach to the dissemination of European Official Statistics.

Cultural differences of European peoples are not solved by means of the use of ELF. National statistical yearbooks are contextualized and addressed to a specific national community, but when they are translated into English they are addressed to an international public which does not share the same cultural background, context and experience. Changing recipients is a challenge for culture-bound statistics and its dissemination. Nunn (2005: 63) notices that:

> 'International' communication seems to require multiple competences. Studies of pragmatic and discourse competences, that focus on the process of achieving mutual intelligibility in whole spoken or written texts, are assuming increasing significance […] Traditionally, however, "communicative competence" has been used to refer to the adaptation to single and well-established speech communities. Preparing for communication between people from a broad range of backgrounds, who will often communicate beyond their own or their interlocutors' speech

> communities in some kind of ill-defined third zone, implies the need to have a highly developed repertoire of communication strategies.

The language used for disseminating statistics to a broad range of different speech communities in the EU is ELF. This aspect, as stressed by Nunn, implies communicative and language strategies oriented to users' needs, and not to the source language or text.

When drafting EU texts, the English language used by drafters has a low degree of cultural marks (Tosi 2007), which makes the text easily translatable into other languages. The same thing does not happen when a national publication is drafted in a national language, because the text is addressed to national readers and is neither conceived for translation nor to address an international audience. Once completed the tanslation, the language of those texts appears to be more similar to Eurostat ELF. This is due to the work of translators who want to meet the readers' needs. However, even in translated National Statistical Yearbooks there are no elements of a new contextualization needed by international readers. When the National Statistical Yearbooks are translated or, as it is in the case of Ireland, the Statistical Yearbook is proposed in English to an International community, there are no additional explanations or examples, which would, instead, be very useful to this purpose. This problem was not only detected by corpus-assisted analysis but also by means of the survey submitted to European Statisticians. EU statisticians did not find NSIs texts for dissemination clear enough, and proposed a review of texts when addressed to the international community. EU statisticians found Eurostat Statistical Yearbooks clearer than National Statistical Yearbooks translated into English. The Eurostat effort to re-contextualize national statistics at European level resulted to promote clarity and accessibility.

The effort of the EU to be by the side of the readers is also manifested by means of personification in EU bodies. The use of *s*-genitive referred to them is interpreted as a way to overcome the bureaucratic face of the EU institutions.

Another aspect which emerged during the study is the different approach of NSIs when dealing with data dissemination out of their national borders. When searching for information in English, the countries who became EU members in recent years have resulted to have a higher degree of compliance with EU dispositions in fostering accessibility by means of English translations. Some countries could not be included in the ENSY corpus, like Germany, France, Spain and Belgium, due to the lack of English available texts. The same happened when proposing the survey to European statisticians of those countries: they did not reply. EU new entries wanted to state they are worth being European. This is not the case with historical EU member countries who have no need for validating their credibility. This can be one of the reasons why countries like Hungary and Bulgaria have their statistical website completely in English and permanently up-dated, differently from Germany or France, where only a few statistical texts are accessible in English.

One more aspect should be taken into account. As already mentioned in chapter 1, French and German languages are considered more international than Hungarian and Bulgarian, but times have changed and many users study English and prefer this language to access European National Statistics.

To conclude, it can be said that there is still much to be done to improve the language of European statistics, but transformation is in progress.

It would be very interesting to keep on studying the language of statistics, enlarging the ENSY corpus following Sinclair's (1991: 25) suggestion:

> […]. It is now possible to create a new kind of corpus, one which has not final extent because, like the language itself, it keeps on developing.

The inclusion in ENSY of more recent Statistical Yearbooks would enable us to study language development in a diachronic comparison. Another aspect for investigation could be the study of a specific social topic (e.g. population ageing) and comparison of the different national publications to understand how statistics presentation influences or is influenced by the national approach to it. Moreover, the study of the language of statistics can highlight cultural differences, and be a support to Eurostat in promoting a stronger European Statistical Culture, removing obstacles in favor of a clearer and more efficient communication.

ANNEX 1

Informative Questionnaire on Language in Statistics

This questionnaire is intended for gathering information on the use of *English in the discourse of statistics*. It is part of a research in *English for Special Purposes*, a PhD programme of the University of Naples *Federico II*. All the information collected will be used exclusively for the above stated purposes.

Date:

Section 1: Respondent data

1. Name: (M) (F)

_____

2. Country of birth:

_____

3. Age .....

a) *under 40* ***b)*** *41-50* ***c)*** *51-60* ***d)*** *above 60*

4. Education level.....

a) *University degree* ***b)*** *Master's degree* ***c)*** *PHD*

5. Field of study

a) *Economics* ***b)****Mathematics* ***c)*** *Statistics* ***d)****Humanities* ***e)****Other (specify).....*

6. First Language/mother tongue:

_____

7. What is your level of English?

a) *Elementary;* ***b)*** *Intermediate;* ***c)*** *Advanced;* ***d)*** *Excellent.*

8. State the International English Language Certification you possess, if any:

_____

9. Where do you work ?....

a) *National Institute of Statistics (country)……..;* ***b)*** *Eurostat;* ***c)*** *University;* ***d)*** *other (specify)……*

10. Which area of statistics do you work in?.....

a) *Social* ***b)*** *Business* ***c)*** *Demographic* ***d)*** *Other(specify).....*

11. Have you ever used English to communicate on statistical topics in the last 5 years? *YES* *NO*

12. If you answered YES specify on which occasions, please choose one or more answers:.....

a) *abstracts;* ***b)*** *papers;* ***c)*** *articles;* ***d)*** *publications;* ***e)*** *readings;* ***f)*** *meetings;*

g) *conferences;* ***h)*** *classes;* ***i)*** *others (specify)……..*

Section 2: Accessibility and Clarity

1. Do you consider that Accessibility and Clarity[46] refer also to the language used to present/explain statistics? *YES NO*

2. For the aim of Accessibility and Clarity which do you consider the most important.....

a) *Table presentation is more relevant than language of texts*

b) *Language of texts is more relevant than table presentation*

c) *Table presentation is as relevant as language of texts*

3. Do you consider that the requirements of Accessibility and Clarity are leading to popularisation/simplification of statistical language ?
YES NO

4. Regulation (EC) N° 223/2009 [47]relates Clarity and Accessibility to Quality. On a scale from 1 to 10 (bottom to top) which rank would you give to the relevance of language used to present/explain with respect to quality in statistics ? _____

5. Do you have any feed back on the English versions of NSIs publications regarding whether they are accessible and clear to users ?
YES NO

6. If you answered YES to question 5, is the feed-back by expert users or non-expert users ?.....

---

[46] ***The European statistics code of Practice -Principle 15 –ACCESSIBILITY AND CLARITY –*** *European statistics should be presented in a clear and understandable form, disseminated in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata and guidance.*

[47] ***Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 – art. 12- Statistical Quality:*** 1.*To guarantee the quality of results, European statistics shall be developed, produced and disseminated on the basis of uniform standards and of harmonised methods. In this respect, the following quality criteria shall apply: (e) 'accessibility' and 'clarity', which refer to the conditions and modalities by which users can obtain, use and interpret data;*

a) *Expert users* ***b)*** *non-expert Users* ***c)****both*

7. Have you ever read the textual parts of Eurostat Yearbook in English?
YES NO

8. If you answered YES to question 7, did you find it clear ?
a)*YES b)ENOUGH c) NOT COMPLETELY d) NO*

9. What would you suggest to improve clarity:

h) *Include a greater number of examples*
i) *Improve glossaries*
j) *Shorten sentences*
k) *Provide additional explanations when referring to typical national phenomena*
l) *Use a different lexicon*
m) *Use more visual representation*
n) *Other (specify).....*

10. Have you ever read the textual parts of the English translation of EU-member-country Statistical Yearbooks? *YES NO*

11.If you answered yes to question 10, did you find the yearbooks texts clear?
a)*YES b) ENOUGH c) NOT COMPLETELY d) NO*

12. Concerning the text, do you consider the English translation of statistical publications in your organization meets the requirements of Clarity and Accessibility to international readers?
YES NO DO NOT KNOW

13. In your opinion what would improve Accessibility and Clarity when addressing international users ? (you can choose more than one answer).....

k) *Produce an original English version in the place of translation*
l) *Include a greater number of examples*
m) *Improve glossaries*
n) *Shorten sentences*
o) *Provide additional explanations when referring to typical national phenomena*
p) *Use a different lexicon*

q) *Choose more expert translators*
r) *Further revision of English texts by Expert statistician with high English knowledge.*
s) *Use more visual representation than in the original text*
t) *Other (specify)……..*

REMARKS and SUGGESTIONS

……
……
……

THANK YOU FOR YOUR TIME !

Bibliographical References

Adab, Beverly (2000). Evaluating Translation Competence. In Shäffner, Christina/Adab, Beverly (eds.), *Developing Translation Competence*. Amsterdam/Philadelphia: John Benjamins, 215-228

Altenberg, Bengt (1982). *The Genitive v. the Of-Construction. A Study of Syntactic Variation in 17th Century English*. Malmo: CWK Gleerup.

Baker, Mona (1995). Corpora in Translation Studies: An Overview and some Suggestions for Future Research. *Target* 7(2): 223-243.

Baker, Mona (1996). Corpus-based Translation Studies: The Challenges that Lie Ahead. In Somers, Harold (ed.), *Terminology, LSP and Translation: Studies in language engineering in honour of Juan C. Sager*. Amsterdam/Philadelphia: John Benjamins, 175-186.

Baker, Mona (ed.) (2001). *Routledge Encyclopedia of Translation Studies*. London / New York: Routledge.

Baker, Mona (2006). *Translation and Conflict*. London-New York: Routledge.

Barber, Charles (1985). Some Measurable Characteristics of Modern English Prose. In Swales, John (ed.). *Episodes in ESP*. Oxford: Pergamon Press, 3-14.

Bernardini, Silvia / Zanettin, Federico (2004). When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In Mauranen, Anna / Kujamäki, Pekka (eds), *Translation Universals: Do they exist?*. Amsterdam/Phladelphia: John Benjamins Publishing Company, 51-62.

Bhatia, Vijay K. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, Douglas / Johannson, Stig / Leech, Geoffrey / Conrad, Susan/ Fineganet, Edward (1999). *Longman Gramman of Spoken and Written English*. Oxford: Longman.

Biber, Douglas (2003). Compressed noun-phrase structure in newspaper discourse: the competing demands of popularization vs. economy. In Aitchison, Jean / Lewis, Diana M. (eds), *New Media Language*. London and New York: Longman, 169-81.

Blum-Kulka, Shoshana (1986).Shifts of cohesion and coherence in translation. In House, Juliane/ Blum-Kulka, Shoshana (eds.) *Interlingual and Intercultural Communication*.Tubingen,Gunter Narr Verlag, pp.17-35

Bowker, Lynne / Pearson, Jennifer (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Byrne, Jody (2006). *Technical Translation, Usability Strategies for Translating Technical Documentation*. The Netherlands: Springer.

Caliendo, Giuditta (2003) The multilingual voices of Europe: The European Commission Translation Service. In Lima, Maria (ed.), *Language Culture and Politics. Issues and Debates in Political Sciences.* Rome: CUEN, 11-20.

Chesterman, Andrew / Gallardo San Salvador, Natividad / Gambier, Yves (eds) (2000). *Translation in Context*. Amsterdam: John Benjamins Publishing.

Chesterman, Andrew (2004). Beyond the particular. In Mauranen, Anna / Kujamäki, Pekka (eds), *Translation Universals: Do they exist?*. Amsterdam/Phladelphia: John Benjamins Publishing Company, 33-49.

Christ, Oliver (1994) A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the 3rd Conference on Computational Lexicography and Text research (COMPLEX 94)*, Budapest: Hungary, 23-32.

Coates, Jennifer (1983). *The Semantics of the Modal Auxiliaries*. London/Camberra: Croom Helm.

Coates, Jennifer (1995). Root and epistemic possibility in English. In Aarts, Bas / Meyer, Charles (eds) *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press, 145-156.

Collins, Peter (2009). *Modals and Quasi-modals in English*. Amsterdam: Rodopi.

Crystal, David (1995). *The Cambridge Encyclopaedia of the English Language*. Cambridge: Cambridge University Press.

Crystal, David / Derek, Davy (1985). *Investigating English Style*. London: Longman.

Crystal, David (2003). *English as a Global Language* (Second edition). Cambridge: Cambridge University Press.

Di Martino, Gabriella / Polese, Vanda (eds) (2005). *'Languaging' and Interculturality in EU Domains.* Napoli: Arte Tipografica Editrice.

Dudley-Evans, Tony / St. John, Maggie Jo (1998). *Development in English for Specific Purposes: A Multi-disciplinary Approach*. Cambridge: Cambridge University Press.

Fairclough, Norman (1992). *Discourse and Social Change*. Cambridge: Polity Press.

Fairclough, Norman / Cortese, Giuseppina / Ardizzone, Patrizia (eds) (2007). *Discourse Analysis and Contemporary Social Change*. Vol. n. 54 in the series *Linguistic Insights, Studies in Language and Communication* Bern: Peter Lang.

Firth, John Rupert (1957). *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Firth, Alan (1996). The discursive accomplishment of normality: On "lingua franca" English and conversation analysis. *Journal of Pragmatics* 26: 237-259.

Flowerdew, John / Gotti, Maurizio (eds) (2006). *Studies in Specialized Discourse*. Bern: Peter Lang.

Gee, James P. (1991). *An Introduction to Discourse Analysis*. London: Routledge.

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

Gotti, Maurizio (2003). *Specialized Discourse: Linguistic Features and Changing Conventions*. Bern: Peter Lang.

Gotti, Maurizio (2005). *Investigating Specialized Discourse*. Bern, Peter Lang

Gotti, Maurizio (2006). *Studies in Specialized Discourse*, M. Gotti / J. Flowerdew (eds). Bern: Peter Lang

Gotti, Maurizio (2007). 'Globalisation and Discursive Changes in Specialised Contexts', in N. Fairclough / G. Cortese / P. Ardizzone (eds) *Discourse and Contemporary Social Change*. Bern: Peter Lang, 2007, 143- 172.

Gotti, Maurizio (2011). *Investigating Specialized Discourse*. 3rd Rev. Edition. Bern: Peter Lang.

Gotti, Maurizio / Dossena, Marina (eds) (2001). *Modality in Specialized Texts. Selected Papers*. Bern: Peter Lang.

Gotti, Maurizio / Heller, Dorothy / Dossena, Marina (eds) (2002). *Conflict and Negotiation in Specialized Texts: Selected Papers of the 2nd CERLIS Conference*. Bern: Peter Lang.

Groefsema, Marjolein (1995). *Can, may, must* and *should*: a Relevance Theoretic account. *Journal of Linguistics* 31: 53-79.

Halliday, Michael A. K. / Hasan, Rukaiya (1976) *Cohesion in English*. London: Longman.

Halliday, Michael A. K. (1988). On the Language of Physical Science. In Ghadessy, Mohsen (ed.) *Registers of Written English: Situational factors and linguistic features*. London: Pinter, 162-177.

Halliday, Michael A.K. (revised by M.I.M. Matthiessen). (2004). *An Introduction to Functional Grammar*. Edward Arnold: London.

Halliday, Michael A. K. 1994. *An introduction to Functional Grammar*, London: Edward Arnold.

Helle V. Dam, Jan Engberg, Heidrun Gerzymisch-Arbogast (2005). *Knowledge Systems and Translation*. The Hague: Mouton de Gruyter.

House, Juliane (1999). 'Misunderstanding in intercultural communication: interactions in English as a lingua franca and the myth of mutual intelligibility'. In Gnutzmann, Claus (ed.), *Teaching and Learning. English as a Global Language*. Tübingen: Stauffenburg, 73-89.

House, Juliane (2003)."English as a lingua franca: A threat to multilingualism?". *Journal of sociolinguistics* 7/4: 556-578.

Hoye, Leo (1997). *Adverbs and Modality in English*. London: Longman.

Huddleston, Rodney D. (1971). *The Sentence in Written English*. Cambridge: Cambridge University Press.

Hülmbauer, Cornelia / Böhringer, Heiker / Seidlhofer, Barbara (2008). Introducing English as a lingua franca (ELF): precursor and partner in intercultural communication. *Synergies Europe* n. 3: 25-36.

Hundt, Marianne / Christian, Mair (1999). "Agile" and "Uptight" Genres: The Corpus-based Approach to Language Change in Progress". *International Journal of Corpus Linguistics* [4:2] :221-242.

Kastberg, Peter (2007). Knowledge Communication: The emergence of a third order discipline. In Villiger, Claudia / Gerzymisch-Arbogast, Heidrun (eds), *Kommunikation in Bewegung: Multimedialer und multilingualer Wissenstransfer in der Experten-Laien-Kommunikation. Festschrift für Annely Rothkegel*. Berlin: Lang, 7-24.

Klinge, Alex (1993). The English modal auxiliaries: from lexical semantics to utterance interpretation. *Journal of Linguistics* 29: 315-357.

Kreyer, Rolf (2003). "Genitive and of-construction in modern written English. Processability and human involvement". *International Journal of Corpus Linguistics* [8:2] :169-207.

Jenkins, Jennifer (2003). *World Englishes*. London: Routledge.

Laviosa, Sara / Braithwaite, Ben (1997). Investigating Simplification in an English Comparable Corpus of Newspaper Articles. In Klaudy, Kinga / Kohn, János (eds), *Transferre Necesse Est. Proceedings of the Second International Conference on Current Trends in Studies of Translation and Interpreting* 5-7 September, 1996, Budapest, Hungary. Budapest: Scholastica, 531-540.

Laviosa, Sara (1997). How comparable can 'Comparable Corpora' be? *Target*, 9(2): 289-319.

Laviosa, Sara (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Lexical Prose. *Meta*, 43(4): 557-570.

Laviosa, Sara (1998). Universals of Translation. In Baker, Mona (ed.), *The Routledge Encyclopaedia of Translation Studies*. London/New York: Routledge, 288-291.

Laviosa, Sara (2002). *Corpus-based Translation Studies: theory, findings, applications*. Amsterdam/New York: Editions Rodopi B.V.

Leech, Geoffrey / Smith, Nick (2006). Recent grammatical change in written English 1961-1992: Some preliminary findings of a comparison of American with British English. In Renouf, Antoniette / Kehoe, Andrew (eds), *The changing face of corpus linguistics*. Amsterdam and New York: Rodopi.

Mair, Christian (2006). Inflected genitives are spreading in present-day English, but not necessarily to inanimate nouns. In Mair, Christian (ed.),

Corpora and the history of English: Festschrift für Manfred Markus. Heidelberg: Winter.

Mauranen, Anna / Kujamäki, Pekka (eds) (2004). *Translation Universals. Do they exist?* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mauranen, Anna (2006a). A rich domain of ELF – the ELFA Corpus of Academic Discourse. *Nordic Journal of English Studies* 5, 2: 145-158.

Mauranen, Anna (2006b). Speaking the Discipline: Discourse and Socialisation in ELF and L1 English. In Hyland, Ken / Bondi, Marina (eds), *Academic Discourse across Disciplines.* Bern: Peter Lang, 271-294.

Meunier, Fanny and Granger, Sylviane (eds) (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins Publishing Company.

Murphy, Amanda C. (2008). *Editing Specialized texts in English. A Corpus-assisted Analysis*. Milano: Edizioni Universitarie di Lettere Economia Diritto.

Myers-Scotton, Carol (2002). *Contact Linguistics: Bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press.

Newmark, Peter (1988). *La traduzione: problemi e metodi*. Garzanti Editori.

Nunn, Roger (2005). Competence and teaching English as an International Language. *Asian EFL Journal*, September 7 (3): 62-74.

Nyerere, Julius (1990) cited in Mussa Lupatu, Daily News of February 9, 1990 -Appendix No. 1/24

Olohan, Maeve (2004). *Introducing Corpora in Translation Studies*. London: Routledge.

Ostler, Nicholas (2010). *The Last Lingua Franca. English until the Return of Babel*. New York: Walker Publishing Company.

Palmer, Frank R. (1990). *Modality and the English Modals*. London: Longman.

Pistillo, M. Giovanna (2005). *Two Languages, Two Cultures, Two Worlds: the Interpreter's Challenge*. Tesi di Dottorato in *Lingua Inglese per Scopi Speciali* (ESP), Università degli Studi di Napoli Federico II.

Perkins, Michael (1983). *Modal Expressions in English*. London: Frances Pinter.

Piga, Antonio (2011). Hybridization and change in the Discourse of the European Commission. In Sarangi, Srikant / Polese, Vanda / Caliendo, Giuditta (eds), *"Genre(s) on the Move. Hybridization and Discourse Change in Specialized Communication.* Napoli: Edizioni Scientifiche Italiane, 99-120.

Pym, Anthony (2004). On the Pragmatics of Translating Multilingual Texts. *The Journal of Specialised Translation* Issue 1 January 2004: 14-25.

Sandrini, Peter (2006). Translation and Globalization. In Gotti, Maurizio / Šarčević, Susan (eds), *Insights into Specialized Translation*. Bern: Peter Lang. 107-120.

Schubert, Klaus (2011). Specialized Communication Studies: An Expanding Discipline. In Petersen, Margarethe / Engberg, Jan (eds), *Current Issues in LSP Research: Aims and Methods* Bern: Peter Lang.

Seidhlhofer, Barbara (2005). Key concepts in ELT: English as a lingua franca. *ELT Journal* Vol. 59/4 October 2005: 339-341.

Seidlhofer, Barbara (2006). English as a lingua franca in the expanding circle: What it isn't. In Rubdy, Rani / Saraceni, Mario (eds), *English in the World: Global Rules, Global Roles*. London: Continuum, 40-50.

Seidlhofer, Barbara / Breiteneder, Angelika / Pitzl, Marie-Luise (2006). English as a Lingua Franca in Europe: Challenges for Applied Linguistics. *Annual Review of Applied Linguistics* (2006) 26: 3-34.

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

Snell-Hornby, Mary (1992). The professional translator of tomorrow: Language specialist or all-round expert?. In Dollerup, Cay / Loddegaard, Anne (eds) *Teaching Translation and Interpreting: Training, Talent and Experience*. Philadelphia/Amsterdam: John Benjamins publishing Company, 9-22.

Snell-Hornby, Mary (2000). Communication in the Global Village. In Shaffner, Christina (ed.), *Translation in the Global Village*. Clevedon: Multilingual Matters, 11-28.

Somers, Harold (ed.) (1996). *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*. Philadelphia/ Amsterdam: Benjamins.

Swales, John M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Taviano, Stefania (2010). *Translating English as a Lingua Franca*. Firenze: Le Monnier Università.

Tosi, Arturo (2007). *Un italiano per l'Europa. La traduzione come prova di vitalità*. Roma: Carocci.

Toury, Gideon (1995). *Descriptive Translation Studies – and Beyond*. Amsterdam: John Benjamins.

Trosborg, Anna (ed.) (2000). *Analysing Professional Genres*. Amsterdam: John Benjamins.

Vanderauwera, Ria (1985). *Duch Novels Translated into English: The transformation of a minority literature*. Amsterdam: Rodopi.

Wales, Katie (1996). *Personal Pronouns in Present-day English*. Cambridge: Cambridge University Press.

Wilss, Wolfram (1999). *Translating and Interpreting in the Twenty-first Century*. Amsterdam: Benjamins.

Widdowson, Henry G. (1979). *Explorations in Applied Linguistics.* Oxford: Oxford University Press.

On-line references

Collins, Peter (2007). Can and may: Monosemy or polysemy? *Annual Meeting of the Australian Linguistic Society (7-9 July, 2006)*. In Mushi, Ilana / Laughren Mary (eds), St Lucia, Australia: School of English, Media & Art History, University of Queensland. Available at: http://espace.library.uq.edu.au/eserv/UQ:12785/Collins_ALS2006.pdf (Last accessed 7 February 2013)

European Commission, Directorate General for Translation (2008). *Translating Tools and Workflow.* Available at: http://bookshop.europa.eu/en/translation-tools-and-workflow-pbHC3212080/?CatalogCategoryID=luYKABst3IwAAAEjxJEY4e5L (Last accessed 7 February 2013)

Gardner, Jessica (2009). Making sense of Statistics. Available at: http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/1242.pdf (Last accessed 10 December 2012)

Hinrichs, Lars / Szmrecsanyi, Benedikt (2007). Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 437-474. Available at: http://www.benszm.net/omnibuslit/Szmrecsanyi_Hinrichs_proofs.pdf (Last accessed 7 February 2013)

Jenkins, Jennifer / Seidlhofer, Barbara (2001). Bringing Europe's lingua franca into the classroom. *The Guardian Weekly*, 19 April 2001. Available at: http://www.guardian.co.uk/education/2001/apr/19/languages.highereducation1 (Last accessed 7 February 2013)

Kankaanranta, Anne (2009). Business English Lingua Franca in intercultural (business) communication. Available at: http://www.languageatwork.eu/downloads/LAW%20Business_English_Lingua_Franca_in_intercultural_business_communication.pdf (Last accessed 2 January 2013)

Laviosa, Sara (1996). The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation. Available at: http://www.llc.manchester.ac.uk/ctis/phd/completed_phd/laviosa/ (Last accessed 7 May 2012)

Kastberg, Peter (2008). Cultural Issues Facing the Technical Translator. *The Journal of Specialised Translation.* Available at: http://www.jostrans.org/issue08/art_kastberg.pdf (Last accessed 3 November 2012)

UNECE – *Making Data Meaningful* (2006, 2009, 2011). Available at: http://www.unece.org/fileadmin/DAM/stats/documents/writing (Last accessed 15 April 2012)

OECD – Glossary. Available at: http://stats.oecd.org/glossary/detail.asp?ID=3764 (Last accessed 19 October 2011)

OECD – *Innovative Approaches to Turning Statistics into Knowledge* (2008). Available at: www.oecd.org/oecdworldforum/statknowledge (Last accessed 15 April 2012)

Statistics Denmark (2003). *Good Dissemination Practice in Statistics New Zealand and Statistics Denmark.* Available at: http://www.dst.dk/pukora/epub/upload/6841/gooddissem.pdf (Last accessed 15 April 2012)

Official websites of the EU- member-state NSIs and Eurostat

(Last accessed 15 April 2012)

http://www.mof.gov.cy/mof/cystat/statistics.nsf/index_en/index_en?OpenDocument
http://www.dst.dk/en
www.statistik.at/web_en/
http://statbel.fgov.be/en/statistics/figures/
http://www.stat.ee/en
http://www.stat.fi/index_en.html
http://www.insee.fr/en/default.asp
https://www.destatis.de/EN/Homepage.html;jsessionid=1B00A9F668E59ED43BA80D27F9A948AC.cae2
http://www.statistics.gr/portal/page/portal/ESYE

http://www.cso.ie/en/
http://www.istat.it/en/
http://www.stat.gov.lt/en/
http://www.csb.gov.lv/en
http://www.statistiques.public.lu/en/actors/statec/index.html
http://www.nso.gov.mt/
http://www.cbs.nl/en-GB/menu/home/default.htm
http://www.stat.gov.pl/gus/index_ENG_HTML.htm
http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_main&xlang=en
http://www.statistics.gov.uk/
http://www.czso.cz/eng/redakce.nsf/i/home
http://portal.statistics.sk/showdoc.do?docid=359
http://www.stat.si/eng/index.asp
http://www.ine.es/en/welcome_en.htm
http://www.scb.se/default____2154.aspx
http://www.ksh.hu/?lang=en
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/