Research meets practice: t-survey 2005

An online survey on terminology extraction and terminology management

Daniel Zielinski
(d.zielinski@mx.uni-saarland.de)

Yamile Ramírez Safar
(ramirez.yamile@googlemail.com)

Linguistic Data Processing Section
Department 4.6 - Applied Linguistics & Translating/Interpreting
Saarland University

Abstract

This paper reports the results of an online survey on terminology management and terminology extraction conducted by the Linguistic Data Processing section of Dept. 4.6 - Applied Linguistics & Translating/Interpreting at Saarland University. The survey was available on the Internet in English, French, German, and Spanish from mid-May until September 2005 [1]. It was promoted in many major CAT mailing lists, and by translator and interpreter associations (BDÜ, ITI). Over 400 professional translators, terminologists, and interpreters all over the world have responded to the questionnaire.

With this survey we want to investigate the relationship between research and practice in the area of terminology extraction and evaluate if there is any need to reconcile both. Aimed at translators, terminologists, interpreters and project managers, the main goals of the survey are to investigate the dissemination and application of terminology management tools (with a focus on terminology extraction tools) and to assess the demands on today's terminology extraction tools.

1. Introduction

Terminology management has now been taught for several years at universities and in translation-oriented education programmes. It continues to be a lively research field involving applied linguists, computational linguists, and computer scientists.

One of the most active research areas in terminology management in the last fifteen years has been *terminology extraction* (TE). A surprising amount of applied research in this field has produced several different approaches (linguistic, statistical, and hybrid) as well as tools to facilitate the time-consuming process of terminology compilation. In recent years, there has been an increasing demand for terminology extraction software (c.f. e.g. Warburton, 2001). However, although a number of term extractors are already available on the market today, it seems that they are still not widely used nor do they meet the real needs of translators, interpreters, and terminologists.
Today, there is no doubt about the increasing importance of terminology in our society. Terminology plays a very important role in many different fields such as standardization, translation, technical documentation, and software localization. Accurate and complete terminology improves the productivity of translators and technical writers, and is a prerequisite for successful communication. Consequently, terminology management dealing with the preparation, processing, and documenting of a specialist vocabulary has become an increasingly important activity.

In the past 25 years, methods, applications and tools for professional terminology work have changed and developed substantially. Despite these improvements, the creation, organization, and management of terminology are still time consuming and cost-intensive tasks. As has been pointed out by the LISA Terminology Survey Report (2001), term identification and extraction are the tasks that are the most

frequently performed manually and that would significantly benefit from automation.

Terminology extraction can be defined as the operation of identifying term candidates (TC) in a given text and should be terminologically distinguished from term recognition, which is the operation of comparing the TC lists (the output of TE) with a terminological database (TDB) in order to identify known/unknown terms.

According to Lieske (2002), the term extraction process comprises four different tasks, namely a) the compilation of a corpus (machine readable texts, that serve as a basis for the TE), b) terminology extraction (identification and extraction of TCs), c) the evaluation or validation of the results (determining the terminological relevance of a TC), and finally d) the classification of terminology according to classes and categories.

TE has three major applications: terminology management, translation, and information retrieval (Thurmair 2003), the first and second being the focus of our investigation. In terminology management and translation, TE is used for the creation, enlargement, and update of terminology. The main goals are to facilitate the task of identifying terms and their translations and to increase the productivity of translators, interpreters, and terminologists. With respect to the number of languages involved in the extraction process, we can basically distinguish between *monolingual* and *bilingual* TE. In the translation process, monolingual TE is usually carried out before the translation starts, either on the source texts or on reference texts. The goal is to identify all relevant terminology in a text to be translated or - in the case of terminology - to identify all terms of a certain (sub-) domain. Bilingual TE, on the other hand, is mostly carried out on translated texts (parallel corpora or translation memories). Its main goal is the recognition of potentially equivalent terminological units in two languages (Thurmair, 2003). In both cases the resulting TCs are often compared to existing terminological databases in order to distinguish known terms from unknown terms (cf. Saß 2004).

2. Approaches to terminology extraction

Depending upon the type of information used as a basis to identify terms, approaches to TE are usually classified as linguistic, statistical or hybrid. Linguistic and statistical approaches can be further subdivided into term-based (intrinsic) and context-based (extrinsic) methods (cf. Bourigault et al. 2001, Streiter et al. 2003).

Terminology extraction tools (TETs) following a linguistic approach try to identify terms by their linguistic structure, e.g. morphological and syntactic structure. For this purpose, texts are annotated with linguistic information with the help of morphological analysers, part-of-speech taggers and parsers. Then TCs with a certain tag structure are filtered from the annotated text by using pattern matching techniques. Intrinsic methods try to filter TCs according to their internal structure, e.g. according to the morphological structure (e.g. German "*Zylinderabschaltung* (cylinder cutout)" ds=zylinder#ab_$schalten~ung). Extrinsic methods try to identify TCs by analysing the morpho-syntactic structure of a word or phrase, such as looking for part-of-speech sequences like NP= noun + noun (e.g. printer menu). Another technique is to filter TCs by looking for commonly used text structures such as definitions and explanatory contexts like "X is defined as " or "X is composed of " (cf. Pearson 1998, Saß 2004).

The assumption underlying the statistical approaches to TE is that specialized documents are characterized by the repeated use of certain lexical units or morpho-syntactic constructions. TETs based on statistics try to filter out words and phrases that appear with a frequency higher than a given threshold by applying different statistical measures (see Manning & Schütze 1999 for an overview). Term-based statistical methods try to compute the structure of TCs e.g. by using n-grams. Often, the structure of existing terms is compared to the structure of words and phrases in a corpus in order to filter out TCs with a similar n-gram structure (example-based approach; cf. Streiter et al. 2003). Another common method is to compare the frequency of words and phrases in a specialized text to their frequency in general language texts assuming that terms tend to appear more often in specialized texts than in general language texts.

Different evaluation criteria exist for TETs, involving among others accuracy, supported file formats, languages, etc. (for a detailed list see Lieske 2002, Sauron 2002, Saß 2004). The most frequently used criteria are noise and silence, and recall and precision. While noise refers to the ratio between discarded TCs and the accepted ones, silence refers to the number of terms not detected by a TET. Recall and precision are two measures frequently used in IR, the former being defined as the ratio between the sum of correctly retrieved terms and the sum of existing terms, the latter being defined as the ratio between correctly extracted terms and the sum of proposed TCs (c.f. Zielinski 2002).

In order to eliminate noise from the resulting TC lists, linguistic and statistical approaches often make use of stop word lists. Stop word lists contain "empty" words (cf. Carstensen et al. 2001) that are not of interest for the terminologist, because they do not qualify as terminological units, but that are often filtered as TCs because of their morpho-syntactical structure or because of their frequency value (e.g. articles, pronouns, auxiliary verbs, and prepositions).

Both approaches, linguistic and statistical, have their advantages and disadvantages. TETs following a purely linguistic approach tend to produce too many irrelevant TCs (noise), whereas those following a purely statistical approach tend to miss TCs that appear with a low frequency value (silence; cf. Clematide 2003). Linguistic-based TETs often provide better delimited TCs than statistical-based ones. Furthermore they usually reduce TCs to their canonical form (base form or lemma) and thus do not provide result lists with repeated TCs as is the case for statistical-based tools. However, the disadvantage of linguistically based TETs is that they are language-dependent and thus only available for major languages. Statistical TETs, on the other hand, can also be used for lesser-used languages that lack computational resources such as minority languages (cf. Streiter et al. 2003).

On their own, linguistic and statistical methods still fail to tackle many of the basic inherent problems of TE, such as the reduction of TCs to their canonical form, detection of multi-word terms, term variations, and terms in discontinuous units (i.e. coordinations, juxtapositions). Since the nature of these problems is so varied, it seems that only a combination of approaches will help to provide efficient TETs. In fact, the only method which has been "recovered" by many authors because of its still unexplored possibilities is the *hybrid* approach. Authors such as Cabré (2001) claim that the only way to make progress in the field of automatic terminology extraction is by combining statistical and linguistic methods. Fortunately, this seems to be the approach current investigations in the area of TE are taking (Thurmair, 2003).

Although there are many investigations about the important role semantic information plays in the performance of TETs (e.g. to find out the semantic relations between terms), most TETs currently available on the market do not incorporate a semantic component (Thurmair, 2003). So, from a practical point of view, it can be stated that a *term* is normally considered by TETs as a word or phrase with a particular structure and belonging to a specific category (noun or nominal phrase) which additionally occurs with relatively high frequency (Thurmair, 2003).

3. Terminology extraction software - a short market overview

In this section we will give a short overview of TETs that are commercially available on the market. We will briefly present the major tools and compare their general features and functions, and classify them according to the approaches outlined in the previous section. Tools developed in research, which are in most cases prototypes of limited applicability and dissemination, will not be taken into account. The interested reader is invited to refer to Cabré et al. (2001) for an overview of such systems developed in the 1990s.

As has been the case for terminology management systems, in the beginning many different TETs came onto the market. Many companies tried to offer some term extraction functionality, either integrated in their CAT-tools (e.g. Trados TermExtract, now MultiTerm Extract) or as standalone versions (e.g. TerminologyExtractor from Chamblon). Today it seems that only a few major tools have asserted themselves on the market. Among these figure MultiTerm Extract (Trados), SDL PhraseFinder, Xerox Termfinder, Terminology Wizard (Synthema), and TerminologyExtractor (Chamblon).

Depending upon the underlying technology (statistical or linguistic), supported languages, file formats, license type (freelance, network) etc., the price for TETs varies from around 100 to several thousand euro.

All TETs provide lists of term candidates that can either be exported for external validation (e.g. in *.txt, *.csv) or validated directly in the TET. During the validation process, the user can usually edit TCs, add manually extracted terms, set status flags (e.g. checked, unchecked), and add TCs to stop word or abbreviation lists. While the first TETs only provided the user with very long TC lists with nearly no additional information that could help to validate the TCs, today's TETs usually link TCs to the context, mostly in form of KWIC concordance windows. Linguistically based tools usually also provide information about POS, and canonical form of the TC. Furthermore, all TETs incorporate sorting functions. TCs can be sorted either according to frequency or alphabetical or - as in the case of SDLPhraseFinder 2005 - according to a "confidence index" in which TCs are ranked according to frequency in combination with the term pattern type. Another feature that is intended to speed up the validation process and that has been added recently to TETs is term recognition. In this step TCs are compared to selected terminological databases in order to distinguish known from unknown words.
In order to allow the user the adaptation of the extraction algorithms, most systems provide additional parameters such as the minimum or maximum length of TCs, minimum frequency of TCs, and minimum or maximum of proposed translations in bilingual mode. Trados MultiTerm Extract, for example, gives the user the possibility to adapt the noise/silence ratio by the use of a slider. After the validation process the accepted TCs can either be exported into one of the standard exchange formats or - in the case of integrated TETs - into the respective terminological database. The table below gives a comprehensive overview over some of the major TETs and their properties.

	Trados MultiTerm Extract 7	**SDL Phrase Finder 2005**	**Multitrans (Corpus builder module)**	**Xerox TermFinder**	**Synthema Terminology Wizard 3.0**	**Chamblon Terminology Extractor 3.0**
Term extraction	mono- & bilingual	mono- & bilingual	mono- & bilingual	mono- & bilingual	mono- & bilingual	monolingual
Approach	statistical	hybrid	statistical	hybrid (XELDA Technology)	hybrid	statistical
Supported languages	137 SL & TL	nl, de, en, fr, it, pt, es	all	de, fr, it, en, sv, es, ru, pt, no, hu, fi, nl, da (total 15 languages)	SL= de, en, fr, it, pt, TL=SL+af, da, ca, nl, pl, pt, ru, es, cz, tr, hr	en, fr, + ?
Term identification (single/multi word terms)	SWU, MWU	SWU, MWU	SWU, MWU	SWU, MWU	SWU, MWU	SWU, MWU
Parameters	Min_length, max_length, max_num, QA-filter (slider), search for new translations, max_trans, min_trans_freq	Exclude_uppercase, max_length	Min_length, max_length, list possible translations	no documentation available	Handle unknown words as nouns, convert all terms to lower case, minimal frequency	Context width, frequency filter, minimum frequency, text filter,
Input	One or more	One or more	One or more	One or more	One or more	One or more
File formats	doc, html, jsp, asp, aspx, xml, sgml, qsc, xtg, ttg, tag, rtf, tmx, tmw, ttx, bif, txt	rtf, html, txt, SDL TM, SDL itd	doc, WordPerfect, txt, Unicode text files, html, xml, pdf, TMX-compliant translation memories, etc.	html, sgml, xml, txt, doc, rtf	rtf, html, txt, Trados TM, Transit	doc, html, rtf, txt
Stop lists	yes	yes	yes	yes	yes	yes
Term recognition (existing TDBs)	yes	no documentation available	yes	no	yes	no documentation available
Results	no lemmatization	no documentation available	no lemmatization		lemmatized	
Additional information	context, frequency,	POS, root form,	context, frequency, source,	frequency, context	context	frequency, context
Validation of TCs	editable list, different status codes, (check, delete, add translation, generate sample sentences, add manually extracted terms,	editable list comments,	editable list	editable list	editable list, different status codes, add to dictionary, stop list, abbreviation list	no

	etc.					
Sorting of TCs	frequency, alphabetical	frequency, alphabetical, confidence index	frequency, alphabetical,	no documentation available	frequency, alphabetical	frequency, alphabetical
Concordance window	yes (add as context sentence)	yes	yes	yes	yes	yes
	Trados MultiTerm Extract 7	**SDL Phrase Finder 2005**	**Multitrans (Corpus builder module)**	**Xerox TermFinder**	**Synthema Terminology Wizard 3.0**	**Chamblon Terminology Extractor 3.0**
Export format	MultiTerm, MultiTerm XML, txt,	SDL TermBase online, tab delimited	no documentation available	OLIF II	eif, xls	txt
Exportable info	user definable	user definable	no documentation available	no documentation available	lemmatized term, translation, frequency, status, example sentence or segment	terms, example sentences (not both at the same time), with/Without frequency
Export filter	yes	yes	yes	no documentation available	yes	yes
Component or single product	part of Trados CAT-Tools, integrates with MultiTerm, Workbench, Workbench	integrates with SDL family of products, SDLX, TermBase online	integrates with Multitrans	part of Xerox Terminology Suite, integrates with TermOrganizer	standalone	standalone
Price	Freelance (1055)	?	568.00 - 2060.00	negotiable	500 (mono-), 800 (bilingual)	~ 100

Table 1: TETs and their properties

It should be added that several evaluations of TETs have been conducted by researchers and practitioners. Even if it is not our intention to discuss them here, we would like to refer the interested reader to some of these evaluations that may serve as a starting point for further reading: Saß (2004), Lieske (2002), Sauron (2002), and the report of the *Schweizer Bundeskanzlei* (2001).

4. t-survey 2005 - Description of the survey's design and goals

t-survey was made available on the Internet in English, French, German and Spanish from mid-May 2005 until September 2005. Its target group includes translators, interpreters, terminologists and project managers who work on a freelance basis or as employees. Therefore, in order to reach as many language professionals as possible, the survey was promoted in many major CAT mailing lists, through several terminology portals (TermNet, DTT), and by translator and interpreter associations (BDÜ, ITI). Over 400 professionals all over the world have responded to the questionnaire.

Unlike other terminology surveys carried out before (e.g. the *LISA terminology survey* (2001)), the purpose of our survey was to examine the relationship between research and practice in the area of terminology extraction and to evaluate if there is any need to reconcile the two. In comparison with *t-survey*, the *LISA terminology survey* had completely different purposes, and its research area was differently defined. The Localization Industry Standards Associations (LISA), responsible for providing practical advice, business guidelines and standards information for translation and localization workflow tools, designed a series of surveys in the area of terminology management. These surveys were carried out in 2001 and aimed mainly at the investigation of the methods, tools and practices for managing terminology in the localization industry (Warburton, 2001:1-5).

Our t-*survey* s main goals are to investigate the dissemination and application of terminology management tools with particular focus on terminology extraction tools (TETs), and to assess the demands on today s TETs. Analysis of the users responses should allow us to determine whether there are differences between the requirements of various professional groups working with terminology, i.e. between translators and terminologists, and, if so, what these differences are. Furthermore, it should enable us to summarize the required functionalities and to formulate design criteria for TETs that fit the different user profiles. In summary, the purpose is to obtain useful information which properly used will have an impact both on research as well as on software development.

The survey was designed in a way that no previous knowledge about terminology extraction was

required by the users. For maximum user-friendliness and minimum frustration that may result from lengthy forms with partly irrelevant questions, the forms were designed to change dynamically according to the answers. The survey s design was based on a task model of translation where, for example, the translation of documents was viewed as a combination of tasks to include analysis, terminology research, translation, term extraction, and term management. The total time estimated to answer the questions was approximately 10 minutes.

t-survey is divided into five sections. The first section, called *Personal Information*, is aimed at identifying the users profiles, such as by age, profession, type of employment, professional experience, and type of education/formation.
The second section, *Working Environment*, focuses on the working conditions and resources that professionals use. This refers, for example, to information such as company size (number of employees), available computer facilities (type of operating systems) and availability of Internet resources (type of Internet connection).

Working languages, areas of specialization (such as technology, law, economics or others), text sorts (e.g. operating instructions, patents, software documentation, commercial reports, web pages, business correspondence, laws, certificates, scientific publications, among others), text size (calculated in standard pages) and file formats (*.txt, *.html, *.doc, *.xls, *.pdf, etc.) are some of the types of information requested in the survey s third section called *Translation*, the main goal of which is to gather information about the translation process itself.

In the fourth section, *Terminology Management*, the central point of interest deals with the process of compilation and storage of terminology. Questions are related to the types of terminology work (e.g. monolingual, bilingual, multilingual) and the frequency in which it is carried out, the method of managing terminology (index cards, Word/Excel tables, relational databases, terminology management systems), the types of information that are stored in the terminological data base (for example, nouns, noun phrases, verbs, collocations, etc.), the criteria for recognizing terms, additional information gathered together with terms (such as administrative information, contexts, definitions, grammatical and semantic information, graphics, etc.) and the estimated time to search terminology during the process of translation.

The fifth and last section, *Terminology Extraction*, is aimed at obtaining concrete information on the use of TETs by the respondents, such as the type of tools used or tested (e.g. MultiTerm Extract (Trados), SDLPhraseFinder (SDL), TermFinder (Xerox), TermFinder (Acrolinx) and Autoterm (IAI)), the purpose of carrying out terminology extraction (extraction of terminology from source texts, from reference texts, from translation memories), the performance of the TETs tested or used, and finally the desired functions for TETs and the amount of money available to be spent on a TET with the desired functions included.

Following the design of the survey and viewing translation of documents as a combination of tasks, as previously mentioned, it can be deduced that sections 3-5 of the survey cover three stages of the translation process, namely text analysis, terminology identification and terminology management.
Concerning the specific goals of this survey, we primarily want to identify the profiles of those participants who carry out terminology work and those who do not. The idea was to have information from primary sources about the use of TETs in practice and the reception these tools have among language professionals.

With regard to those participants who claimed to be involved in terminology activities, we are especially interested in the following points:

(1) to what professional groups do they belong,
(2) what are their qualifications,
(3) what is their knowledge level regarding computer technology and terminology,
(4) what are the frequency and type of their terminology work,
(5) for what purposes do they perform terminology work,
(6) what importance is given to terminology tasks,
(7) with what languages and language combinations do these professionals work most of the time,

and
(8) what TETs are mostly used and preferred.

With regard to those participants who do not carry out terminology work, we also want to determine their user profiles and the reasons they give for not taking into consideration terminology tasks in their work. Moreover, we consider it important to learn if these participants have tested some TETs before and if they would be willing to use them, provided that the tools would satisfy their needs. Finally, since the economic aspect was also considered to be important, all participants were asked to indicate the price they would be willing and able to pay for a TET which satisfies their needs.

Based on this detailed overview of the users profiles it should be possible to accurately determine the users requirements, which should be the *leitmotif* for research in the field of TE for the development of TETs.

5. Results and analysis

We will now present the results of the survey and their respective analysis section by section. A total of 451 participants filled in the survey, expending an average time of 15 minutes [2]. The high number of participants is a sign of the great interest translators, interpreters and terminologists have in issues related to translation and terminology management and developments in the area of CAT tools.

5.1 *Personal information*

Especially translators seemed to be very interested in terminology extraction and terminology extraction tools. This can be concluded from the fact that the majority of the professionals who participated in the survey were translators (371). 81 of the participants were interpreters, 56 project managers and 61 terminologists. Since more than one profession could be selected, 250 participants indicated working exclusively as translators, 5 as interpreters, 12 as project managers and 29 exclusively as terminologists (see figure 1). 86 participants indicated having additional professions, which means that a large number of the participants could be considered highly qualified. To the question regarding the working status, 292 of the total participants indicated working as freelancers, 126 as employees and only 24 both as freelancers and employees (see figure 2).

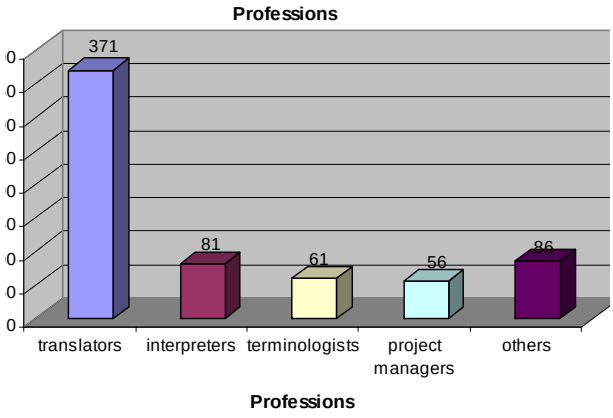

Figure 1

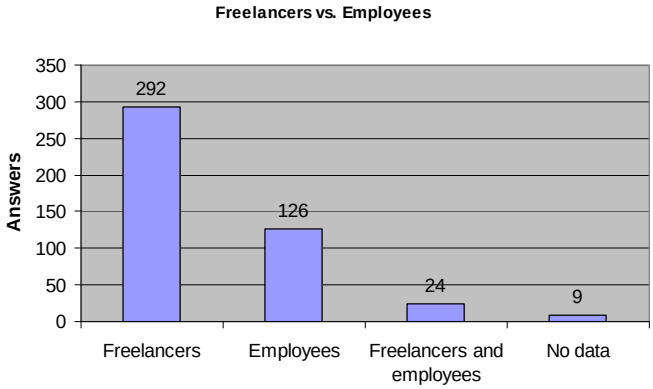

Figure 2

Combining information on professions and working status, it turns out that the majority of translators and interpreters work as freelancers whereas the majority of terminologists and project managers work as employees. Of all professional groups, freelance translators were the largest one (see figure 3).

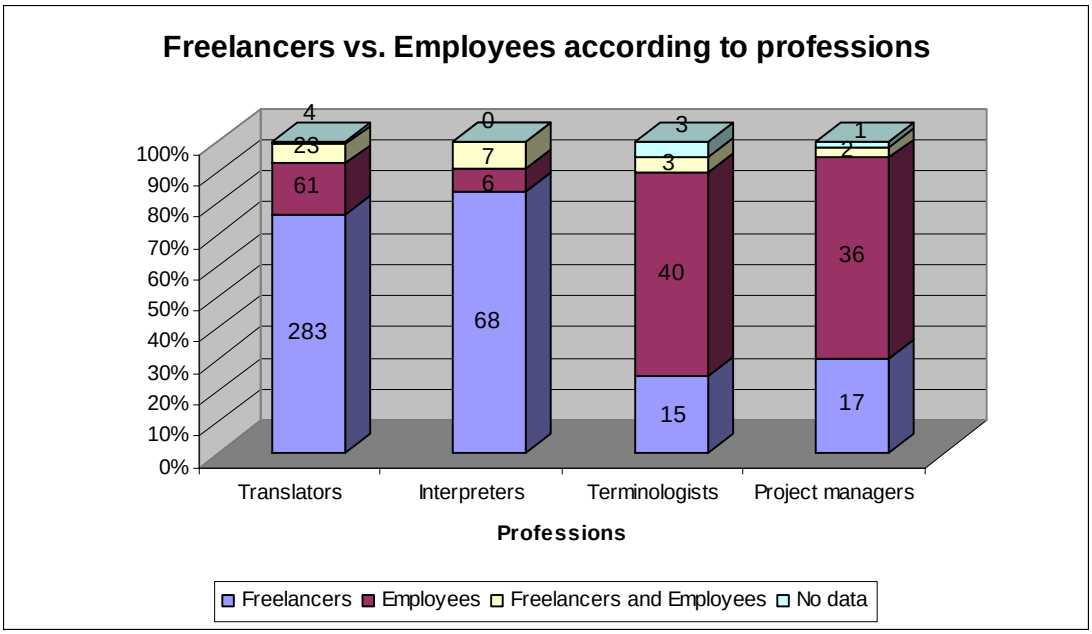

Figure 3

In general, participants were between 20 and 75 years old. For all participants, the average age is around 40 years and the average working experience is 5 years or more (see figure 4).

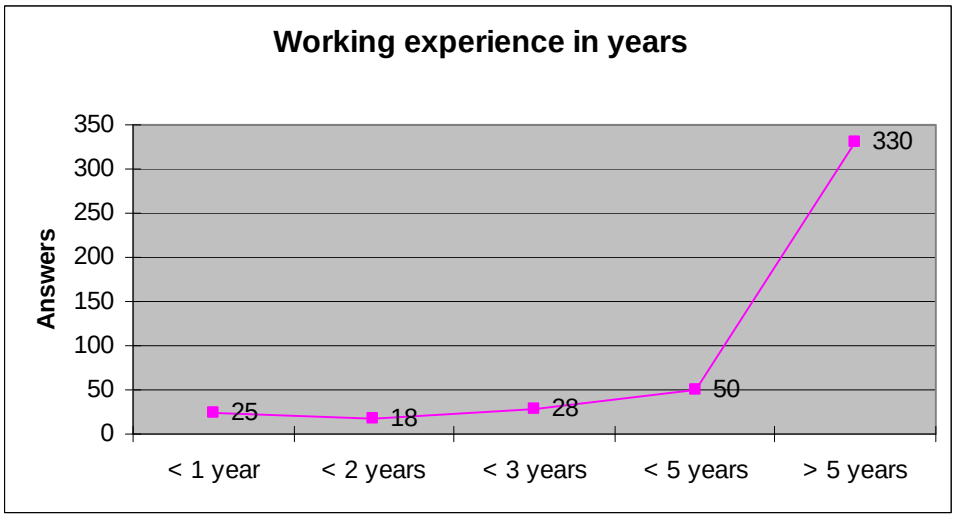

Figure 4

Regarding qualifications, 63% (235) of all translators have a degree as translators; 77% (62) of all interpreters as interpreters; and only 16% (10) of all terminologists as terminologists. 35% (160) of the total participants mentioned other degrees in fields like technology, natural sciences and humanities. Based on these figures, it can be said that the majority of the professionals working as translators and the majority of professionals working as interpreters have a degree in their special field, whereas this does not seem to be the case for terminologists (see figure 5). This may be related to the fact that, although there are universities and educational centres which include terminology courses in their study programs, very few of them offer a degree in terminology.

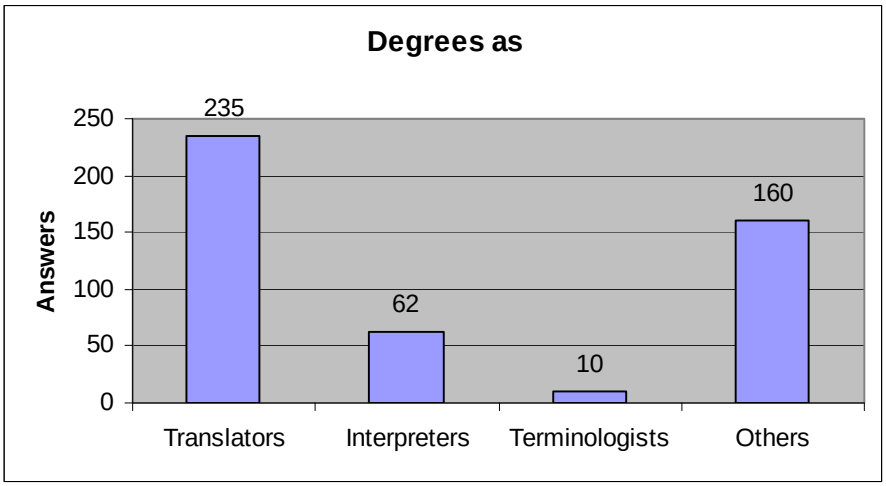

Figure 5

Many participants had additional qualifications such as certifications as translators or interpreters or have attended specialization courses: 29% were certified translators, 43% certified interpreters and 6% both certified translators and interpreters (see figure 6). The number of additional qualifications underlines, as mentioned previously, that the majority of participants are highly qualified.

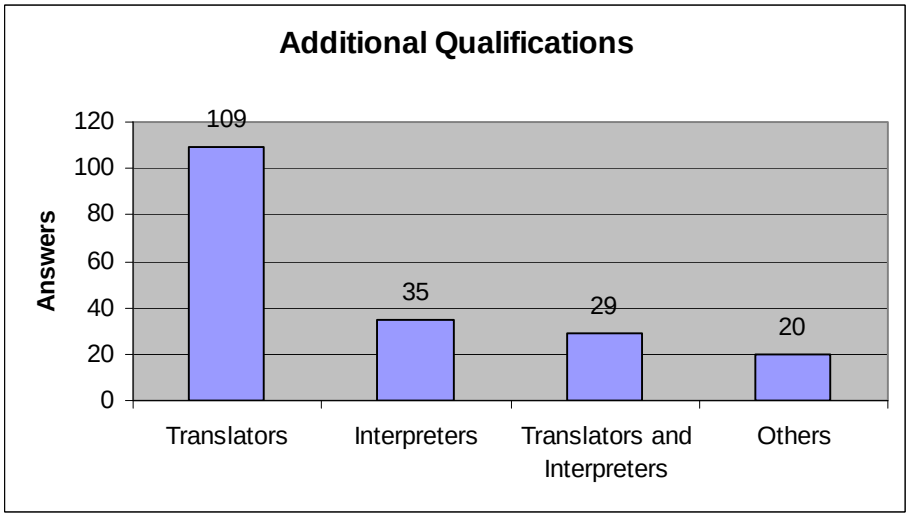

Figure 6

5.2 *Working environment*

61% of all participants answered that they work in companies with 1 to 5 people; 3% in companies with 5 to 10; 4% in companies with 10 to 20; 21% in companies with 20 or more people. The large number of participants working in companies from 1 to 5 people is related to the fact that the majority of the participants being either freelance translators or interpreters. From the figures it can be assumed that the majority of employees work in companies with 20 or more people.

Regarding the use of operating systems, it turned out, as expected, that 85% of all participants work with one of the Microsoft Windows operating systems; 6% with Linux and 5% with Mac. The predominance of Microsoft Windows operating systems is maybe due to the fact that nearly all CAT tools are only available for this platform. As far as the Internet is concerned, 96% of all participants indicated using the Internet for their daily work. From the types of Internet connections, xDSL turned out to be the leader with 72%, followed by ISDN (18%) and modem (10%). 75% of all participants have a flat rate whereas only 25% of them have a connection with download or time limit. These results show that translators, interpreters and terminologists make extensive use of the Internet for their work, be it as a source of

knowledge (information on specialized areas, reference texts), as a source of linguistic information (terminology, usage examples, contexts), or as an absolutely indispensable tool when it comes to project management (reception, processing and delivery of translation orders, etc.). In sum, these figures confirm - as is common knowledge - that the Internet is one of the most important tools for translation and terminology work.

5.3 *Translation*

The results of the third section show that all participants work with a wide variety of languages, mainly the official languages of the European Union, with English, German, French, Spanish and Italian predominating (see figure 7). When interpreting these figures, it is important to keep in mind that our survey was available in four of the predominating languages and that it was announced mainly in European mailing lists and Internet portals.

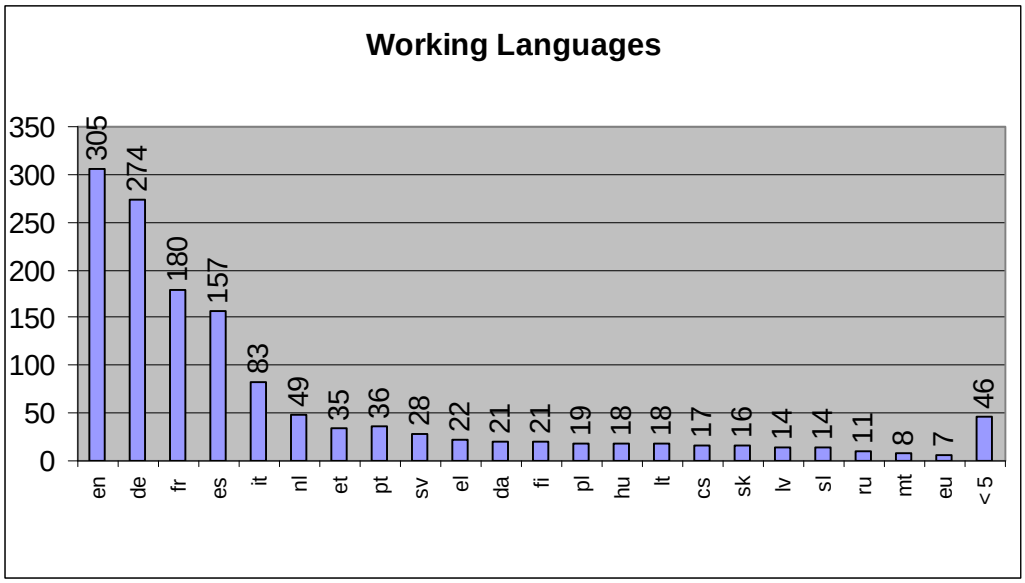

Figure 7

The average number of working languages per participant lies between 3 and 4. These numbers can certainly also be explained by the high percentage of freelance translators who participated in the survey. Concerning the language combinations of the five most frequently used languages, the pair English-German has shown to be the most common one, followed by English-French, German-French, English-Spanish and German-Spanish (see figure 8).

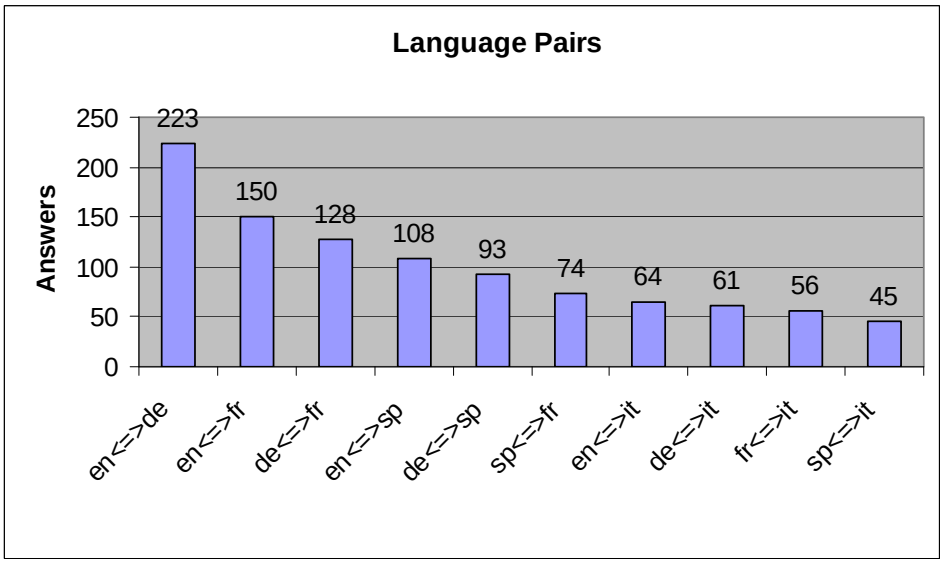

Figure 8

From the areas of specialization, *technology* is shown to be the most important area with 274 answers, followed by *economy* with 222 and *law* with 178 answers. Among other specialization areas mentioned by the participants, *medicine* and *pharmacology*, *arts*, *culture*, *music* and *politics* were the most frequent ones. When comparing the three main specialization areas with the five major working languages, it seems that the proportions between *technology*, *economy* and *law* are equal throughout the languages, which confirms their relevance and predominance in comparison with other areas. It is also worth noting that in all languages *technology* is always the predominating specialization area, followed by *economy* and *law* (see figure 9).

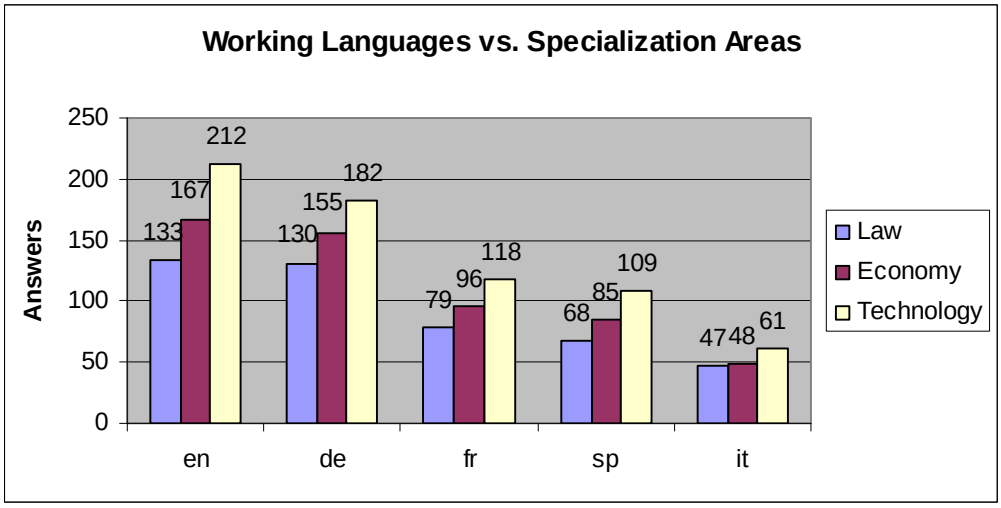

Figure 9

The text types participants most frequently work with are *manuals*, *operating instructions*, *web pages*, *business correspondence*, *software documentation*, *commercial reports*, and *training material* (see figure 10). The text types from *manuals* to *commercial reports* are the ones with the highest demand on translation and all come from the specialization areas *technology* and *economy*, whereas the text types with the lowest demand on translation come from the specialization area *law*. This confirms the results shown in figure 9 where *law* is less frequent than *technology* and *economy*.

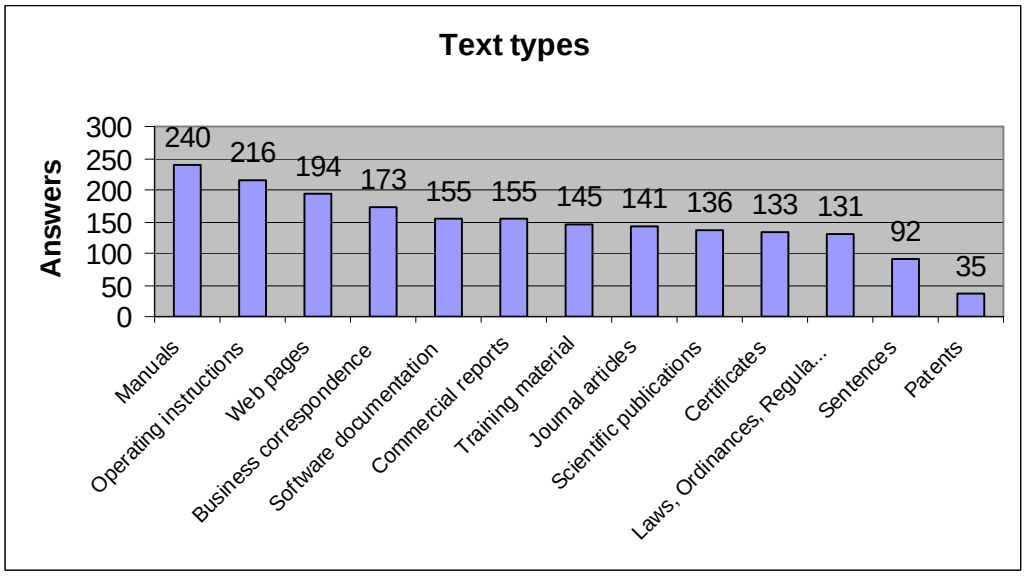

Figure 10

Nearly all participants are used to working with the standard file formats. Since the majority of participants use Microsoft Windows operating systems (as shown in section 5.2), it is not surprising that *.doc is by far the most frequently used file format, followed by *.pdf, *.xls, *.ppt, *.txt, *.html and *.xml. Other formats such as *.rtf, *.fm, *.qxd or *.indd are significantly less frequent (see figure 11). Apart from *.pdf and *.ppt formats, most of the commercial TETs support all these standard file formats (see table 1 in section 3).

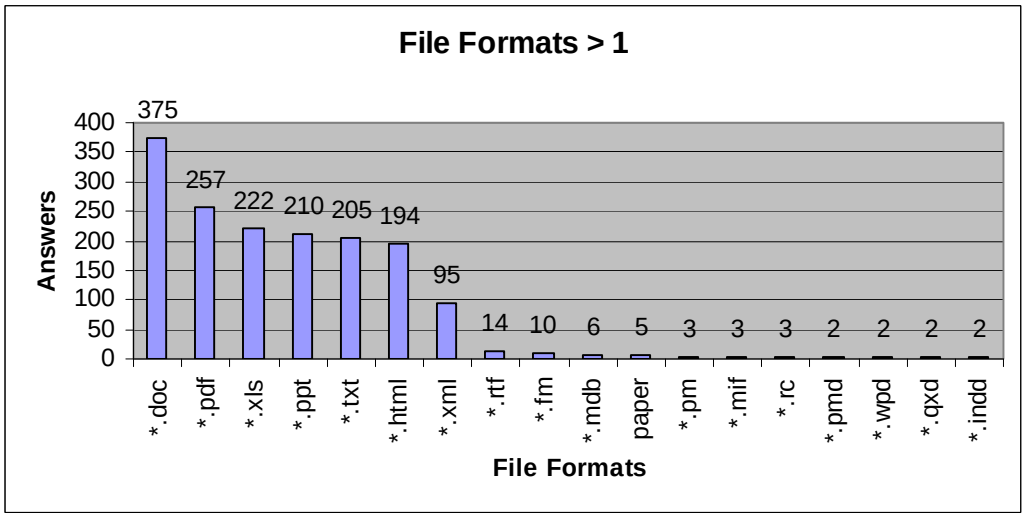

Figure 11

5.4 Terminology Management

91% of all participants carry out terminological work. This is surprising because the majority of participants are freelance translators, and these professionals normally work around the clock, under pressure and have little time left to manage their terminologies. These results, in turn, underline the importance of terminology management that has been outlined in section 1. As can be seen in figure 13 this is true both for translators and interpreters. 45% of the participants indicated that they always carry out terminological work, 45% that they do it only if time permits. Finally, 10% do terminological work only on demand (see figure 12). Participants who do not carry out terminological work said, among others, that terminology work is too time-consuming, that there is no need for it, that terminology is usually delivered by the customer or by a terminology service provider or one s own terminology department.

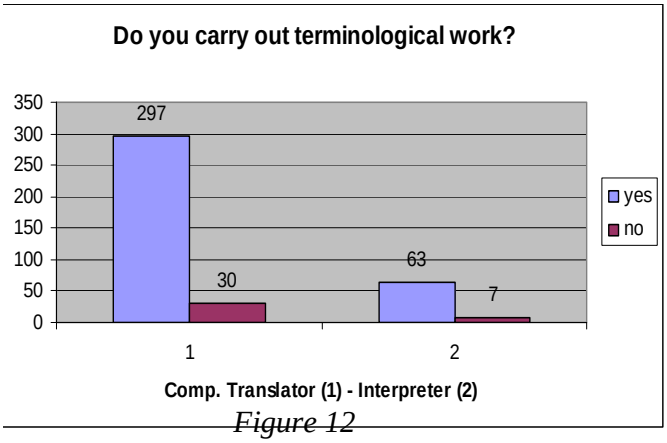

Figure 12

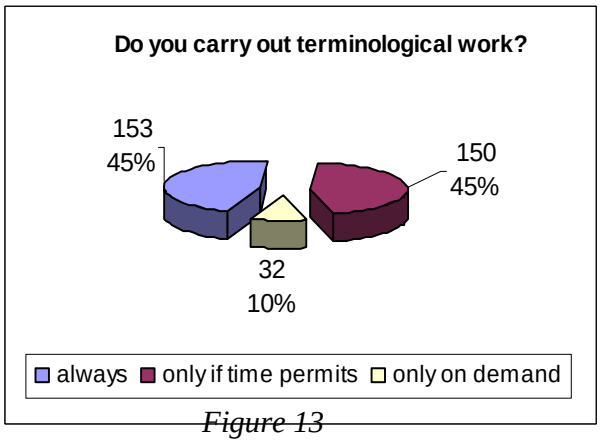

Figure 13

Regarding the type of terminology work, 63% of all participants stated that they perform bilingual terminology work, 23% that they perform multilingual terminology work and 14% that they perform monolingual terminology work (see figure 14).

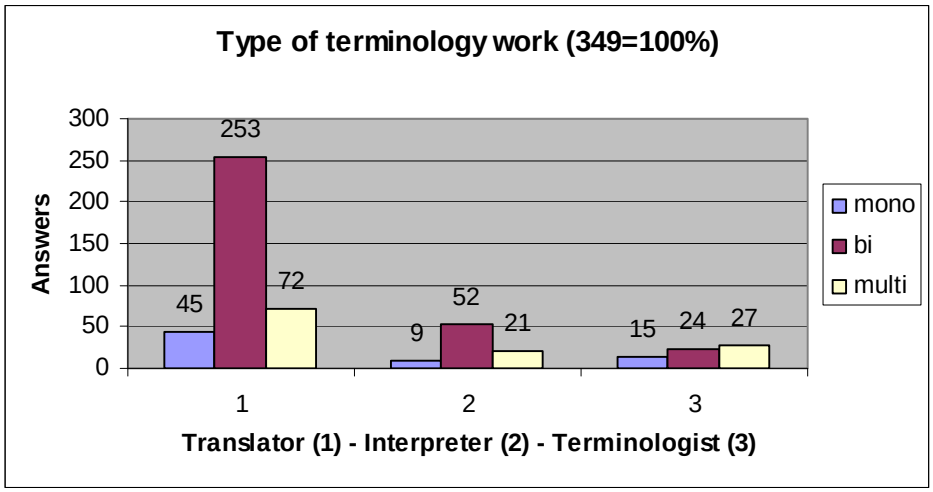

Figure 14

When comparing the different types of terminological work in relation to the professions, for translators and interpreters, the percentage for bilingual terminology work is significantly higher than the percentage for mono- and multilingual terminology work. On the other hand, in the case of terminologists, the percentage of multilingual terminology work is higher than mono- and bilingual terminology work. It is worth noting that for terminologists the percentage of monolingual terminology work is significantly higher than in the case of translators and interpreters. This is probably due to the job profile of terminologists, since their main goal is not translation, but primarily management and standardization of terminology.

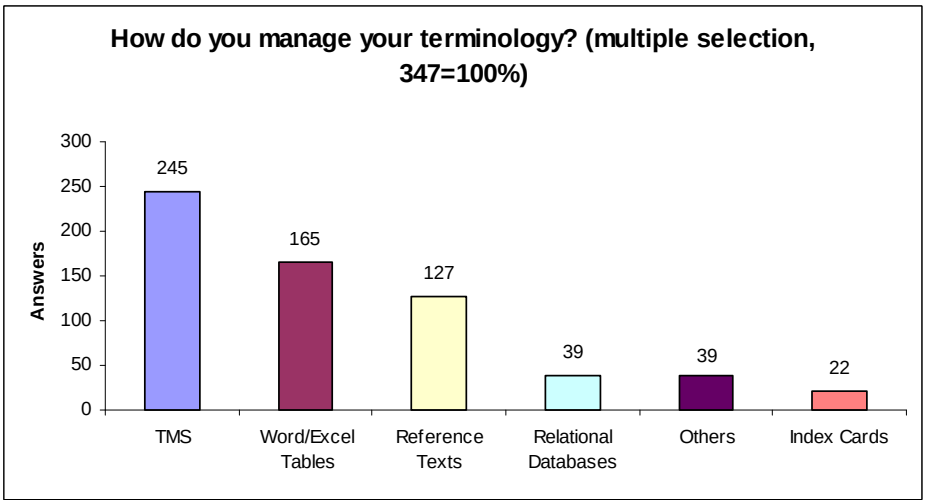

Figure 15

Terminology management is mostly carried out by using appropriate terminology management systems (TMS) (71%). Word and Excel tables (48%) and the storage of reference texts as source of terminology (37%) also play an important role. Relational data bases (10%) and index cards (6%) are far behind (see figure 15). Again this confirms the trend among translators to be up-to-date when it comes to advanced technologies (cf. section 5.1).

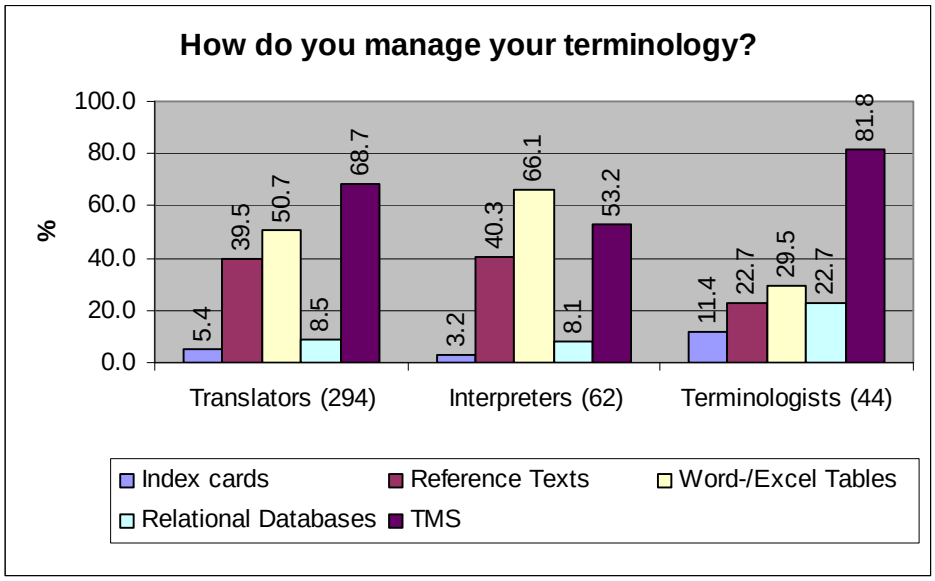

Figure 16

When comparing the use of TMS among the professions, more terminologists (82%) use TMS than translators (69%) and interpreters (53%). Word and Excel tables are used more by translators (51%) and interpreters (66%) than by terminologists (see figure 16). These figures can be explained by the fact that terminologists are very specialized when managing terminology and thus require more specific terminology management solutions. One possible explanation for the relatively high percentage of Word and Excel tables used by translators and interpreters may be that word processor and spreadsheet software have a wider dissemination than TMS, and that translators and interpreters are more familiar with the use of this standard software. Another possible explanation might be that translators and interpreters are not as familiar with the principles and methods of terminology management as terminologists are.

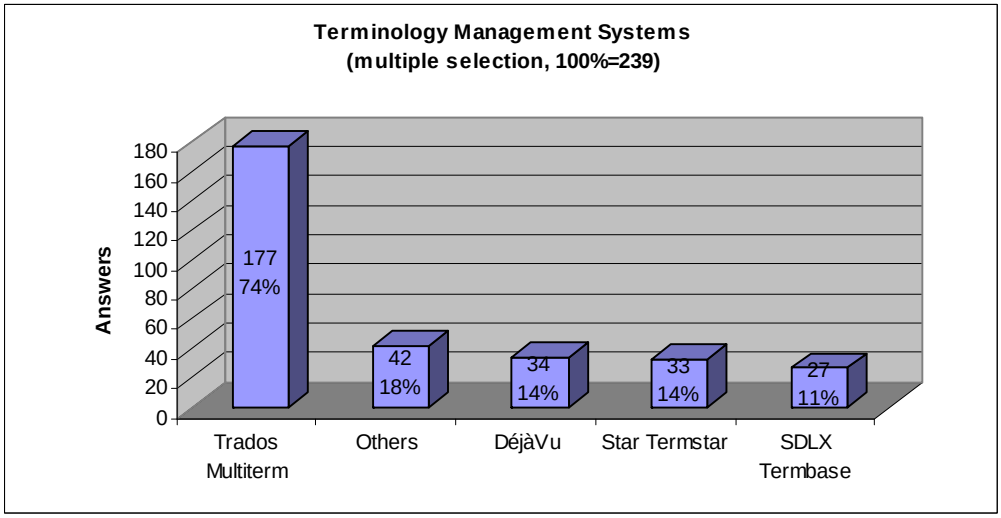

Figure 17

When it comes to TMS, TRADOS Multiterm is the undisputed market leader. 74% of all the participants who answered the question on the use of TMS work with TRADOS Multiterm. Termstar, Déjà Vu and SDLX Termbase are far behind (see figure 17). Some participants mentioned that they carry out terminology work using the glossary and terminology management components of other translation memory systems like Wordfast (8%), Across (1%) and Multitrans (1%). 3% of the respondents work with TMS developed on their own. It is worth noting that the predominance of TRADOS may be, among others, related to the fact that it was one of the first commercial terminology management systems on the market.

Regarding the identification of terminology, translators, interpreters and terminologists indicated that they do not only store nouns and noun phrases (96%) in their terminology databases, but also verbs (82%) and collocations (78%).

When it comes to the criteria used to determine the terminological relevancy of a word or word group, pragmatic criteria predominate clearly. 69% of the participants include unknown words in their terminology (familiarity criteria), 67% consider frequency an identification criterion for terms, 35% also use formal criteria - such as similarity to known terms - and 22% indicate other criteria (see figure 18).

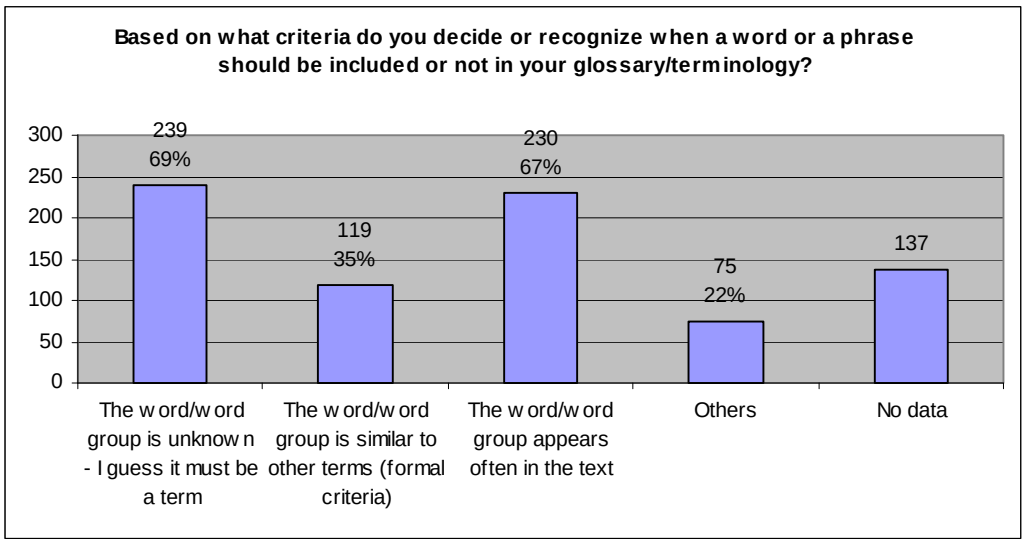

Figure 18

The other term identification criteria indicated by the participants can be clustered as follows: 65% is related to semantics (context, expert knowledge, etc.) and 30% to pragmatics (e.g. time spent on terminology research). 5% of the participants indicated that they identify terms intuitively. The number of participants who did not answer this question at all is surprisingly high (137). This leads us to the assumption that many participants have no conscious but rather intuitive knowledge of how to identify terms.

When asked which type of information they save in their databases, 22% of the participants indicated that they do not save any additional information apart from the term and its translation (see figure 19; nothing). The most frequently saved types of information are contexts (68%; cont), definitions (68%; def), information on usage restrictions (60%; uinfo), subject codes (59%; scode), administrative information such as status, last modified by (40%; adm), grammatical information (33%; gram) and semantic information (synonyms, antonyms) (32%; sem). Less frequently saved types of information include semantic relations (part-whole, hyponym-hypernym) (15%; trel) and figures and graphics (14%; illu). Comparing the types of information stored by translators, interpreters and terminologists, it becomes clear, as expected, that the different information types are more often stored by terminologists than by the other professionals. This can be easily recognized in figure 19, particularly for semantic information, semantic relations, definitions, illustrations, and grammatical information.

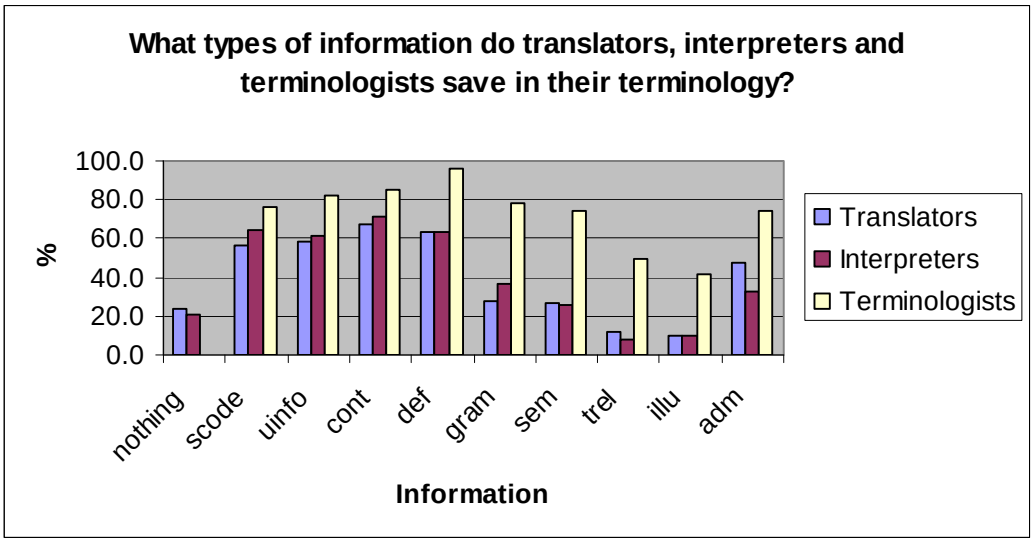

Figure 19

There is a surprisingly high percentage of participants who indicated that they get terminology from their customers (80%), whereas only 1% always get terminology from their customers, 49% sometimes and 50% rarely. 20% do not get any terminology from their customers at all. When asked if the delivered terminology was complete or not, 59% answered 'no', 2% answered 'yes' and 39% answered 'sometimes'. Though the majority of freelance translators said that their customers frequently provide terminology, most of the terminology is incomplete. This means that, in the end, translators definitely have to spend time on terminology work (investigation, validation, etc.).

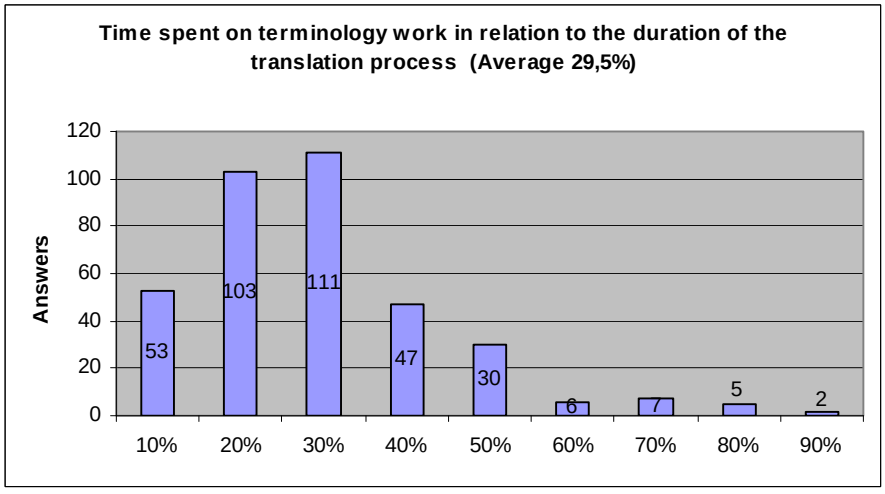

Figure 20

In fact, as indicated by translators and interpreters, the average estimated time expended on terminology work is about 30% of the total time expended on a translation order. This is certainly a significant and by no means negligible amount. When carrying out highly specialized translations, this percentage may reach up to 70% or more (see figure 20). Therefore, automation in terminology management would probably be very welcome by translators and interpreters.

5.5 Terminology Extraction

In order to determine the dissemination of TETs, participants were asked if they have heard of TE before, and if they had any experience with TE tools. While the majority (71%) of all translators, interpreters and terminologists have already heard of terminology extraction, 29% have never heard of it before. Only 15% of all participants that answered the questions in this section use TETs for their daily work; the

majority (85%) do not use them. People who work as terminologists use TETs significantly more frequently (40.9%) than translators and interpreters (15.2% and 13.9% respectively). Nevertheless, the majority of the terminologists still do not use TETs (59.1%) (see figure 21).

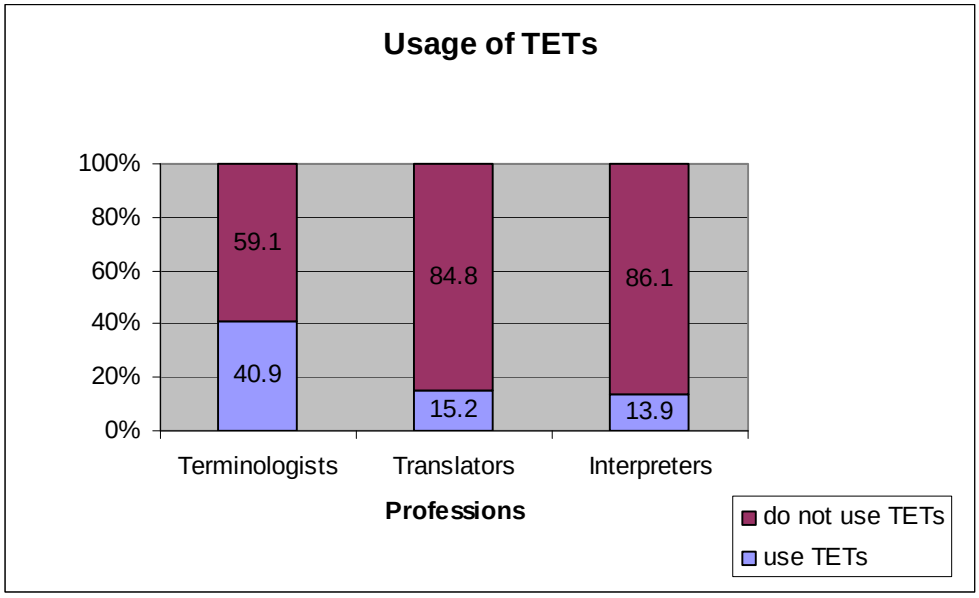

Figure 21

Some of the reasons given by participants for not using TETs were the following in order of importance:

1. The performance of terminology extraction tools is not good enough (98)
2. The results lists of the terminology extraction contain too many wrong term candidates (85)
3. Terminology extraction tools find very few terms (71)
4. The time spent checking the list of term candidates is too long (55)
5. Based on the results lists of the terminology extraction I cannot decide whether a word is a term or not (47)
6. Terminology extraction tools are too expensive (15)
7. I have never heard about terminology extraction tools (9)

TETs are mainly used to extract terminology from source texts, translation memories and reference texts. Figure 22 shows the proportions. The difference between the results regarding terminology extraction from source texts (43) vs. terminology extraction from translation memories (32) can be explained with the working scenario of translators and terminologists. Source texts are always available for translators and terminologists, whereas translation memories often do not exist for specific domains (sub-domains). Reference texts are often unavailable as well and it can be a very time-consuming and difficult task to collect them.

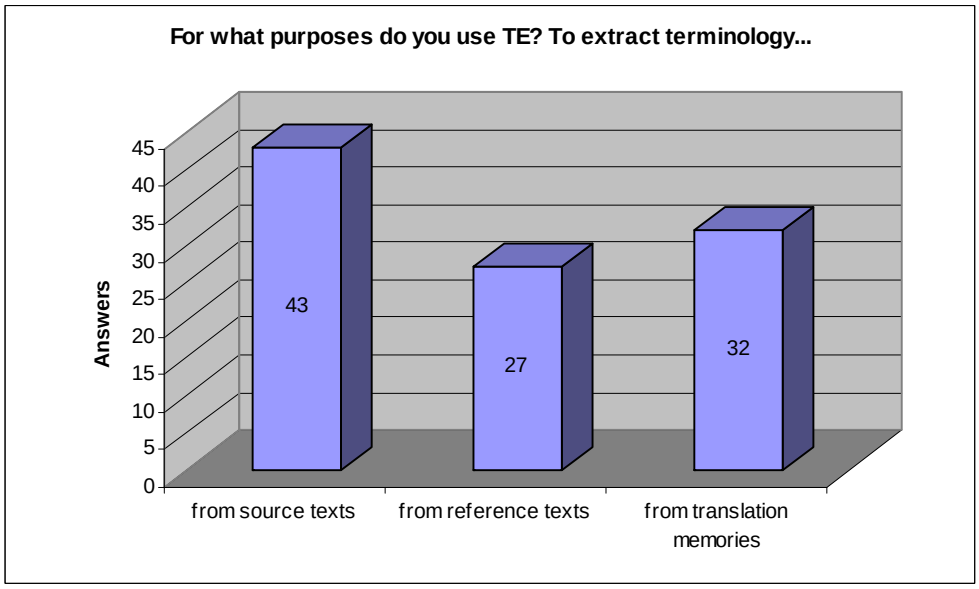

Figure 22

Compared to the actual use of TETs, the participants' potential interest in using these tools is expressed in the following figures: 253 (85%) would use TETs to extract terminology from reference texts, 179 (61%) from translation memories and 169 (57%) from source texts. These figures are interesting, since they are exactly the opposite of the actual usage of TETs. In other words, the participants who work with TETs use them to extract terminology mostly from source texts and less frequently from reference texts, while the potential users of TETs would like to use them to extract terminology primarily from reference texts and secondarily from source texts (see figure 23).

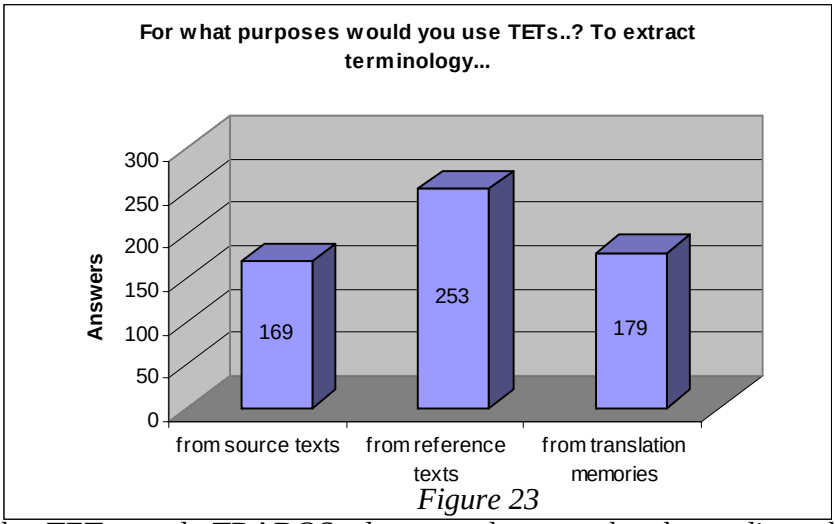

Figure 23

Concerning the TETs used, TRADOS also turned out to be the undisputable market leader with MultiTerm Extract (56%); far behind comes SDLPhraseFinder (12%). The high dissemination of MultiTerm Extract, in comparison with other considered TETs, may be explained by the fact that MultiTerm Extract fully integrates within the most frequently used CAT software package (cf. section 5.4), easily available on the market, and which supports the most file formats and languages (see table 1, section 3). In addition, its price is certainly more affordable when compared to other tools incorporating linguistic knowledge. These tools are significantly more expensive due to their high development costs (see table 1, section 3). Finally, it is also interesting that 12% of the participants use TETs developed on their own and that 7% use concordance programs (e.g. WordSmith) to identify terminology in texts (see figure 24). This confirms - as has been said earlier - that existing TETs still do not meet the users' requirements.

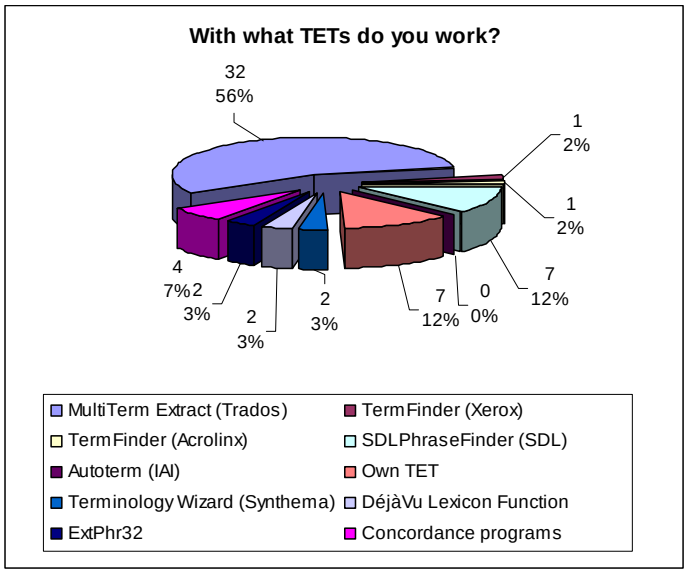

Figure 24

In order to determine the dissemination of TETs, participants who do not use TETs were asked whether they have at least tested them. While 33% of the participants answered that they have tested TETs, 67% answered that they have not. Regarding professions, interpreters who do not work with TETs turned out to be the group which has tested them least (77%), followed by translators (73%). As expected, the percentage of terminologists (73%) that do not use TETs but have tested them is much higher than that of translators (28%) and interpreters (23%). This fact underlines, once again, that terminologists have more specialized needs (see figure 25).

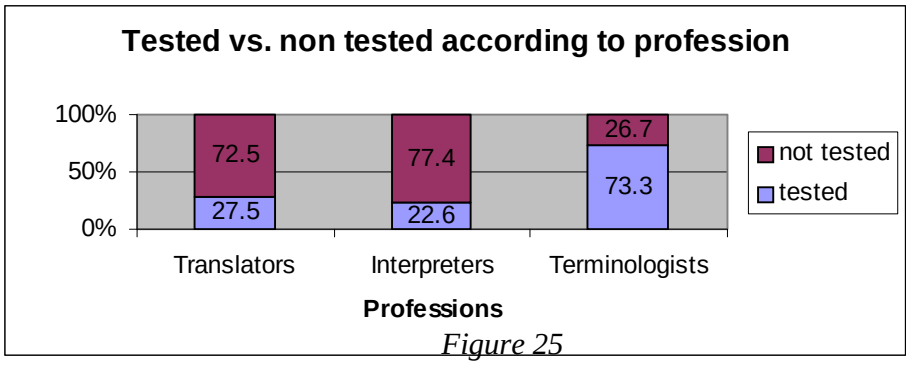

Figure 25

Among the TETs participants have tested, MultiTerm Extract has the highest percentage (55%), followed by others (20%), which interestingly are mostly concordance programs used for TE. This leads us to the supposition that concordance tools might incorporate useful functions for terminology research. 10% of the participants have also tested SDLPhraseFinder, whereas the percentages of other TETs tested (Term Finder (Xerox), Term Finder (Acrolinx), Autoterm), vary between 3% and 4%. This is certainly related to their minor dissemination, which can be explained by their language dependence, higher prices, etc. (see figure 26).

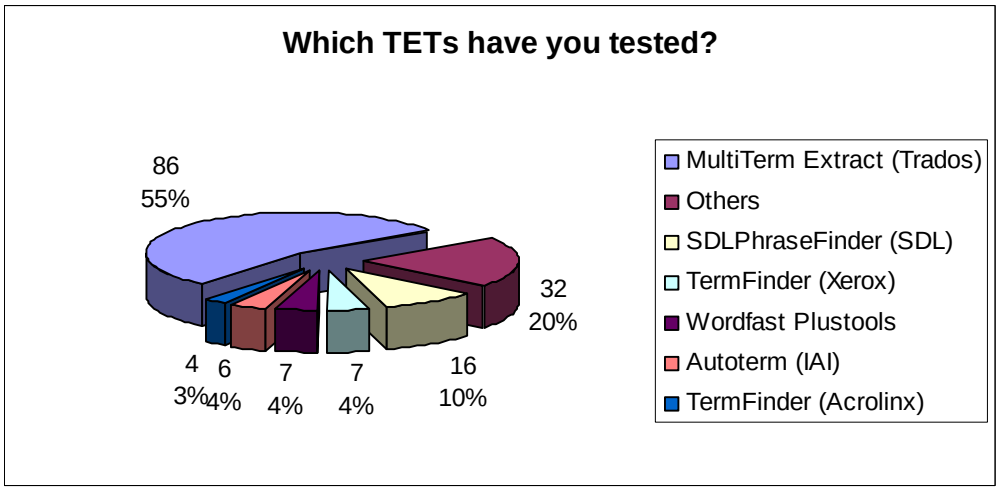

Figure 26

In order to understand which tools tend to be the preferred among the ones tested, we compared the figures of tested TETs with the number of used TETs (see figure 27). Only 30% of those who tested MultiTerm Extract use it for their daily work. The percentage of the people who tested/used the other tools is significantly higher even if these tools are much less disseminated (SDLPhraseFinder 60%, 71% Xerox Term Finder, and 50% Term Finder (Acrolinx) and Autoterm). One possible explanation for these figures might be that language professionals prefer to work with linguistically based term extraction tools instead of statistical-based ones. Unlike MultiTerm Extract, all other TETs include linguistic knowledge.

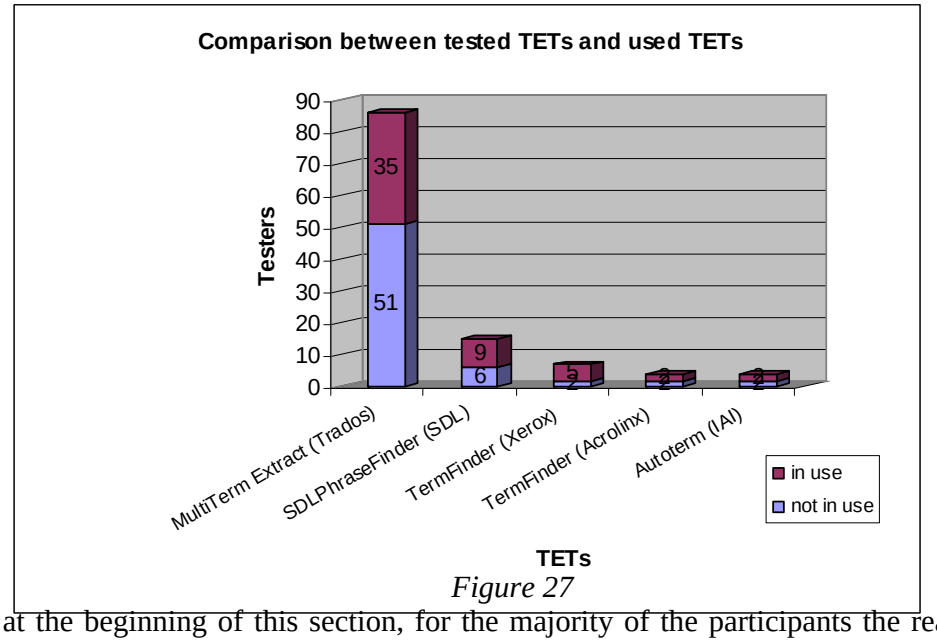

Figure 27

As shown at the beginning of this section, for the majority of the participants the reason for not using TETs is dissatisfaction with the performance of these tools. Since we wanted to make their assessment measurable, we asked the participants to judge the performance of TETs on a scale from 1 (excellent) to 6 (insufficient). According to their answers, the average performance of TETs turned out to have a value of 3.75, which is still far from being excellent (see figure 28).

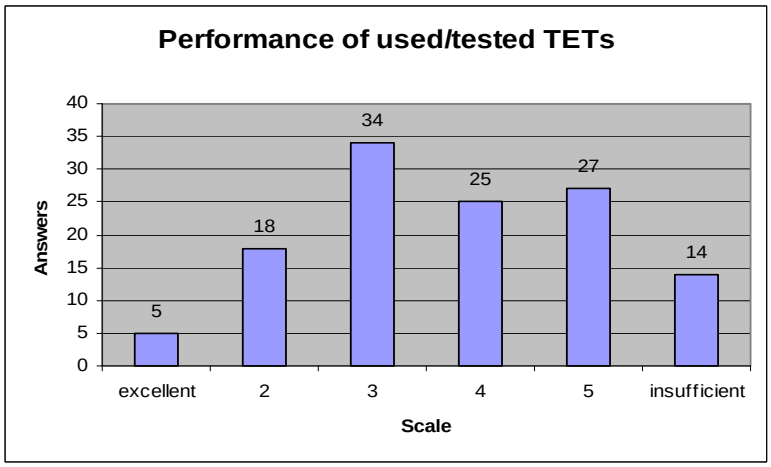

Figure 28

In order to get information about the users' requirements, we asked participants what functionalities they desire in a TET. The majority of the participants (278) want TCs to be linked directly with their contexts. This is probably because context usually plays a very important role in the determination of the terminological relevancy of a lexical unit and thus for the identification of term candidates. However, it is worth noting that only a few participants indicated context as criterion for selecting terms (see section 5.4). This is further support for the assumption that many language professionals identify their terminology intuitively and do not dispose of well defined criteria to identify terms.

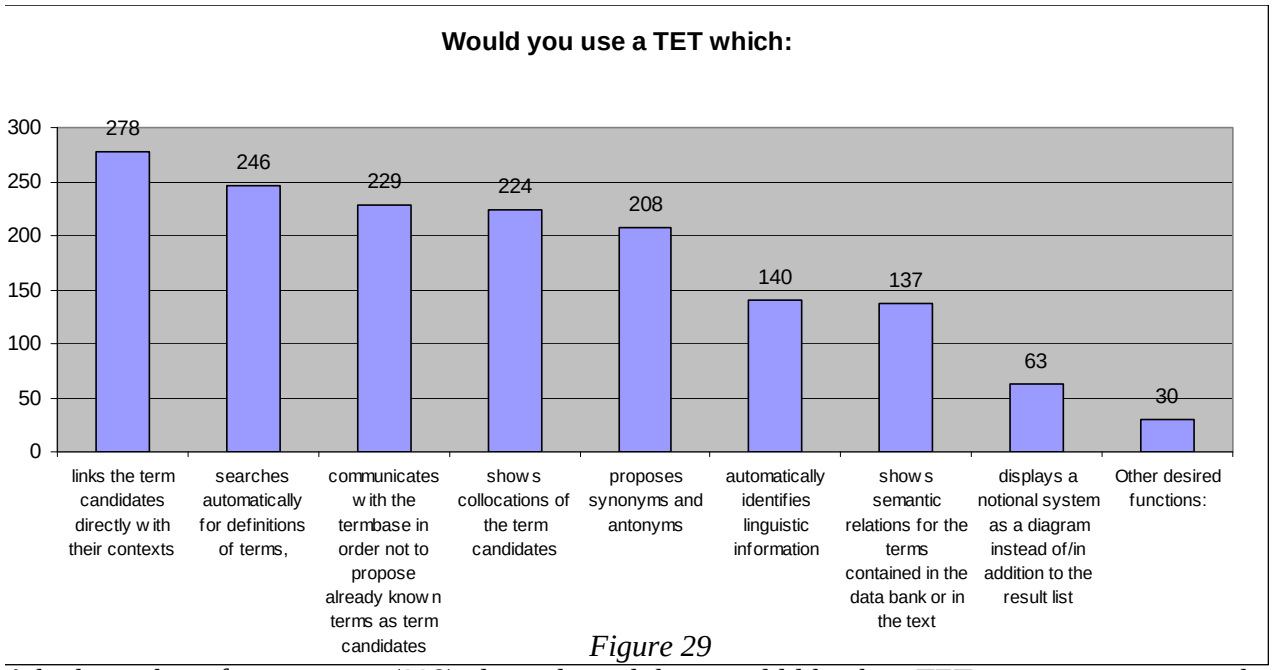

Figure 29

A high number of participants (229) also indicated they would like their TETs to communicate with the terminological database in order not to propose already known terms as TCs, a functionality which is still not available in all TETs (see section 3). Other desired functionalities were, for example, that TETs propose synonyms and antonyms (208), or that they show semantic relations for the terms contained in the database and/or in the text (137), both functionalities clearly in connection with semantic criteria (see figure 29).

Ideas about the price for TETs that provide the functions described above vary from 0 to 5000 according to professions. The majority of the participants (85) are willing to pay between 100 and 250, 54 persons would pay between 250 and 500, 34 between 500 and 750. Only a minority is willing to pay more than 750. The average price that participants would pay for such a TET is 353. One trend is very obvious: the more expensive a TET is, the less probable it is that users would buy it (see figure 30).

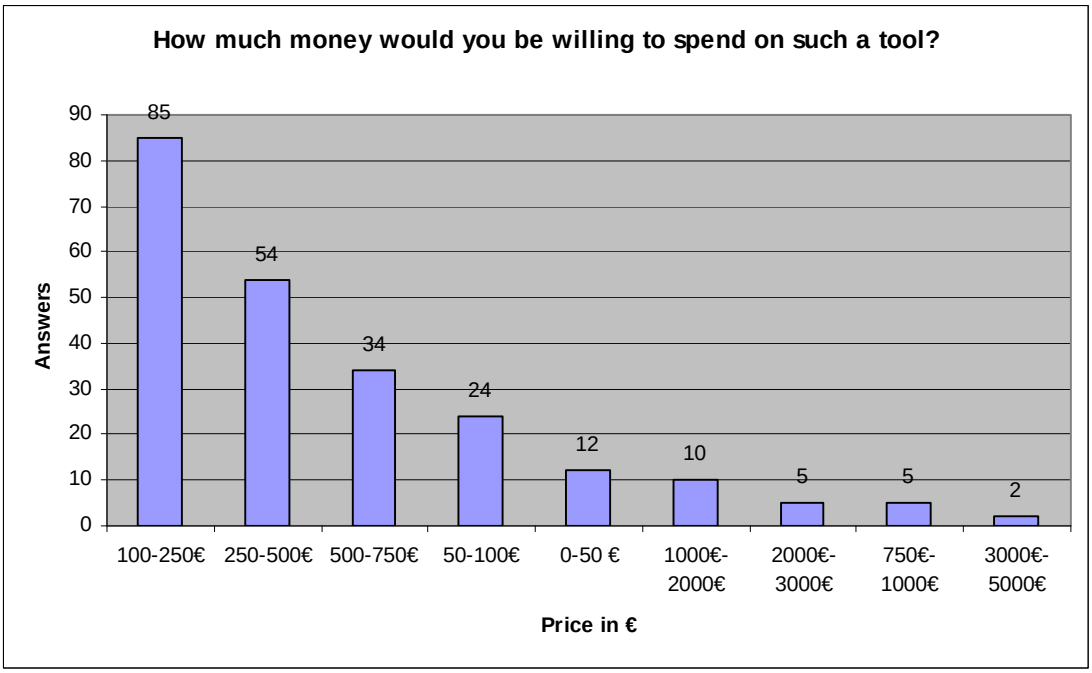

Figure 30

The information on prices is very interesting when compared to the actual prices of TETs available on the market (cf. table 1, section 3). According to table 1 in section 3, Chamblon Terminology Extractor seems to be the only TET to fit the price desired by the majority of the participants (around 100), followed by Synthema Terminology Wizard (between 500 and 800) and MultiTerm Extract from TRADOS (1055 for the freelance edition). The other TETs available are by far more expensive.

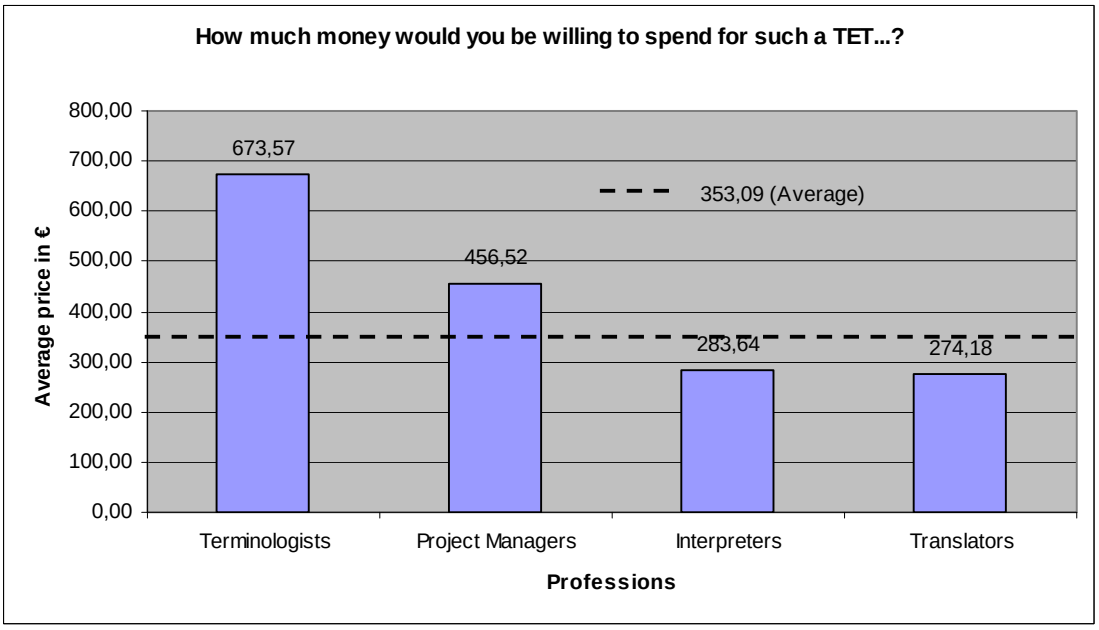

Figure 31

Finally, we compared how much the different professionals would spend on a TET. It turned out that terminologists (674) are willing to pay more than project managers (457), interpreters (284) and translators (274). As mentioned previously, this is certainly due to their more specific needs when working with terminology. For translators and interpreters, these results seem to be coherent since they work mostly freelance and cannot afford to invest much money on a TET. In fact, the prices translators and interpreters would be willing to pay lie under the overall average price expectation (353). In the case

of project managers, the price that these professionals are willing to pay for a TET is higher than the average. This may be due to the fact that they normally work in a company which can probably afford to pay more for CAT tools than freelancers can.

6. Conclusions

The survey showed that even if many translators, interpreters, terminologists and project managers have already heard about terminology extraction, and a relatively large number of them have had some experiences with TETs, TETs are still not widely disseminated. Only few language professionals, mainly terminologists, use TETs for their daily work (15% against 85% that do not use them), which has been explained by their working scenario and their special needs in terminology management.

Based on the survey's results, we know that for many translators and interpreters the basic concepts of terminology work (what is a term, how to recognize it, etc.) are often not clear or systematized. In fact, the majority of professionals apply pragmatic criteria in order to identify terms such as the TC frequency in a text (67%) or the TCs' similarity to other known terms (22%). Concerning terminology management in general, a relatively high percentage of respondents admitted to managing their terminology within CAT systems like WordFast and Déjà Vu, translation memory systems that additionally offer some very basic functionalities for terminology management (mostly glossaries). Regarding TETs, it has been documented that they still do not seem to fulfil the requirements of translators and interpreters, the main reasons being the noise and silence problem, the lack of information for the validation of TCs, and, finally, the high price. The lack of knowledge about the principles of terminology work and - according to the users - the poor performance of TETs could in turn explain why translators and interpreters mostly do not use TETs or instead use translation memory systems for terminology purposes.

Nevertheless, the survey documents that there is a lively interest among all interested professional groups in terminology extraction issues. This interest can be explained by the large part of terminology work involved in translation and the importance of terminology in general. Therefore, many translators, interpreters, and terminologists might be future potential users of TETs.

Translation-oriented studies

Concerning the impact on translation-oriented studies, the results of our survey show that, for example, participants do not seem to be completely sure about the properties of a term and about how terms can be identified. Therefore, they mostly stick to pragmatic criteria to determine if a lexical unit is a term or not. Given the importance of terminology work in the translation process (see 5.4), courses on terminology and terminology management need to become a fixed component in translator and interpreter education programs - which is still not always the case. Furthermore, given that the average age of participants is around 40 years, and that 15 years ago there was nearly no terminology module in translation-oriented study courses, there is also a need for further education courses in terminology work. This should not be understood as merely training courses for the different tools offered by many tools specialists or by the software companies themselves, but rather as courses that focus both on the theoretical and practical side of terminology work and that give an overview of the recent developments in this area. This latter type of course is still very rare.

In sum, the importance of terminology work documents the possibility of an increasing specialization on terminological issues that could result in an increasing need for terminologists and/or terminology service providers. This would advocate for the creation of appropriate terminology-oriented study courses.

For all types of courses that deal with terminology, professional teachers and trainers are still needed. [3]
The same holds true for didactical approaches and methods to train these trainers.

Research

The survey has shown that the majority of participants do not seem to have concrete ideas concerning the semantics of terms. In most cases the semantics of terms was determined intuitively, using structural linguistic knowledge of terms or pragmatic criteria, such frequency. This is, among other reasons, probably due to the lack of detailed semantic descriptions of terms. Therefore, we would suggest that research should further investigate the properties of terms, above all their semantic properties, in order to deliver accurate descriptions of terms that are the prerequisites for the definition of more accurate and practical criteria for term identification.

The results of the survey strongly indicate a trend towards the preference for linguistically-based TETs. These types of TETs require computational resources and tools for high-quality linguistic analysis involving e.g. robust full parsing and semantic tagging. Since such tools are still under development and are not yet applicable in industrial contexts, research still has to make progress in these areas of language technology (see also section 2). This is true for both- major and minor languages.

Another research field besides terminology and computational linguistics that could contribute to the improvement of TETs is *corpus linguistics*. Many participants indicated that they save contexts and definitions as additional information in their terminological databases. Finding useful contexts and definitions is undoubtedly a task that is at least as labour-intensive and time-consuming as term identification itself and thus could also significantly benefit from automation. This requires criteria for determining what makes a context of a term useful, as well as for identifying such contexts. The same is valid for definitions and for techniques for automatic definition extraction.

Software development

Based on the results of the survey it can be determined that many terminologists, translators and interpreters are still not satisfied with the overall performance of TETs and that they demand high-quality tools. One of the direct consequences for the development of TETs is that these tools should be based on linguistic knowledge, implementing language technology in order to fulfil the users' expectations (lemmatized TCs, properly delimited TCs, duplicate TCs, term recognition, term variant recognition, etc.). Since linguistically-based TETs are language dependent, it is obvious that, in the beginning, they can only be developed for the most frequently used working languages. Only if computational resources for minor languages are available can the product palette be gradually extended. In the meantime the only practical solution lies in the (additional) implementation of statistical approaches - knowing that these will not completely fulfil the user requirements. Since translators, interpreters and terminologists mainly work with 3 to 4 languages, the adoption of linguistic approaches to terminology extraction seems justified.

In a few words, the language professionals under investigation do not need tools that function for all languages, but tools that work well for their working languages. Table 2 summarizes the gathered information related to the users' requirements and profiles.

	Terminologists	**Translators**	**Interpreters**
Professional status	employees (66%)	freelancers (76%)	freelancers (84%)
Company size	10-20 (14%), >20 (64%) people	1-5 (76%)	1-5 (80%)
Internet connection	available (flatrate)	available (flatrate)	available (flatrate)
Working languages	2-4	2-5	2-4
Specialization Areas	tech, law, econ	tech, econ, law	law, tech, econ
Text types	web pages, laws, regulations, manuals, scientific	manuals, operating instructions, web pages,	no data
File formats	doc, pdf, html, txt, xls, ppt, xml	doc, pdf, xls, ppt, txt, html, xml	doc, pdf, xls, ppt, txt, html
Term recognition	yes	yes	yes
Integration into CAT systems	yes	yes	yes
Type of TE	multi, bi, mono	bi, multi, mono	bi, multi, mono
Sources for TE	source texts, TM, reference texts	source texts, TM, reference texts	no data

	Terminologists	Translators	Interpreters
Linguistic categories	nouns, verbs, NP/VP phrases	nouns, verbs, NP/VP phrases	nouns, verbs, NP/VP phrases
Term identification	frequency, familiarity, similarity, context	frequency, familiarity, similarity, context	frequency, familiarity, similarity
Required information	cont, def, uinfo, adm, gram, sem, trel, illu	cont, def, uinfo, adm	cont, def, uinfo
Price idea	~ 670	~ 280	~ 280

Table 2: Information related to users' requirements and profiles

One aspect of TETs that has been criticized by participants is the validation process. Many of them indicated that the validation of TCs often is a very time-consuming task: on the one hand, because too many TCs are extracted, and on the other hand, because it is difficult to determine the terminological relevancy of TCs based on the information provided in the results lists. Many participants claimed that additional information such as contexts and definitions are required for TC validation. Again, automatic context extraction and definition extraction could be useful functions. Reducing the time needed for validation seems a necessary prerequisite for the acceptance of TETs.

As has been seen, translators, interpreters and terminologists are not only interested in extracting and saving terms and their equivalents, they also want to gather further information such as contexts, definitions, or additional semantic information (e.g. semantic relations). Therefore, it would be very welcome if the providers of TETs would take a step forward towards the development of such functionalities.

To come back to our initial question, namely how research and practice are related in the area of TE and if there is any need to reconcile both, the results of our survey have shown that research and practice do not completely go separate ways. The existing TETs fulfil the needs of today's users only partially. Theoretically, TETs based on deeper linguistic knowledge, including functions like the ones outlined above, can certainly be developed. We think that a focus on this direction could not only satisfy the needs and expectations of today's users, but also attract new users.

One possibility to discover the exact needs of these potential new users of TETs and to bridge the gaps between research, development, and practice is to encourage and enhance communication between these groups, e.g. through discussion groups and roundtables, or through surveys like the one reported in this paper. In this context we want to refer to the role of terminological networks such as TermNet that help to underline the importance and relevance of terminology work for language professionals, customers, public institutions and industries. They also serve to describe the needs and requirements of language professionals and to make customers, etc. aware of this. Furthermore, they are also a means for informing both groups about recent developments and trends in terminology work. In this way all parties will benefit: First, the industry could develop better targeted tools and consequently sell more, second, translators, interpreters and terminologists would benefit from better tools and consequently increase their productivity. In turn, this is of benefit to the customers.

Finally, we want to underline the importance of the cooperation triangle between *education*, *research* and *practice*. Only a close cooperation among these three areas can guarantee the development of higher quality TETs. First, education in terminology management is fundamental for better-aimed research in terminology related issues. Second, the feedback between research and practice is absolutely indispensable. While research contributes to differentiate between the possible-doable from the not impossible-not doable, practice brings in the already accumulated experience and the economical resources indispensable to determine what is really needed. And third, education and practice contribute to an exchange of ideas and experiences in terminology management to adapt to each others' needs and demands. Only if all these conditions are met can progress towards the optimization of terminology management issues be achieved.

Acknowledgements
We would like to thank Uwe Reinke and Karl-Heinz Freigang of the Linguistic Data Processing Section

of Dept. 4.6 at Saarland University for their comments and help in the different aspects of this survey.

7. References

Bourigault, D., Jacquemin, C. and L'Homme M-Cl, editors. (2001): Recent Advances in Computational Terminology. John Benjamins, Amsterdam.

Cabré Castellvì, M. Teresa; Estopà Bagot, R.; Vivaldi Palatresi, J. (2001): Automatic term detection: A review of current systems. In: Bourigault, D.; Jacquemin, C.; L'Homme, M-Cl, (Editors) (2001), (53-89).

Carstensen, K.-U.; Ebert, C.; Endriss; C.; Jekat, S.; Klabunde, R.; Langer, H. (Hrsg.) (2001): Computerlinguistik und Sprachtechnologie (Eine Einführung). Heidelberg, Berlin: Spektrum Akademischer Verlag.

Clematide, S. (2003): Automatische Termextraktion (TE): Utopie, (kommerzielle) Wirklichkeit, nähere Zukunftt. Gastreferat Nachdiplomkurs Terminologie der ZHW.
Last checked: 28.06.05. URL: http://www.cl.unizh.ch/siclemat/talks/termext03/

Daille, B., Habert, B., Jacquemin, C., and Royauté, J. (1996): Empirical observation of term variations and principles for their description. *Terminology*, 3(2), 197-258.

Heid, U. (2001): Verfahren zur Extraktion von Termkandidaten aus Texten: Ein Überblick. In: Mayer, Felix (Hrsg.): Dolmetschen & Übersetzen Der Beruf im Europa des 21. Jahrhunderts. Freiburg: freigang.maure+reinke, (186-204).

Jacquemin, C.; Bourigault, D. (2003): Term Extraction and Automatic Indexing. In: Mitkov, R. (editor): The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press (599-615)

Lieske, C. (2002): Pragmatische Evaluierung von Werkzeugen für die Term-Extraktion. In: eTerminology - Professionelle Terminologie im Zeitalter des Internet. Akten des Symposiums. Köln, 12.-13. April 2002. Köln: Deutscher Terminologie-Tag e.V. (109-131)

Manning, C.D.; Schütze, H. (1999): Foundations of statistical natural language processing. Cambridge, Londons: MIT Press.

Pearson, J. (1998): Terms in context. Amsterdam: John Benjamins Publishing Company.

Saß, R. (2004): Vergleichende Untersuchung von Terminologie-Extraktions-Tools. Eine computerlinguistische Arbeit mit Englisch und Deutsch. Saarbrücken: Fachrichtung 4.6 - Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen - Universität des Saarlandes (Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen, Band 21)

Sauron, V., 2002: Tearing out the Terms: Evaluating Terms Extractors. Proc. ASLIB 2002, London

Schweizer Bundeskanzlei, Sektion Terminologie / Arbeitsgruppe Extraktoren (2001): Evaluation der Extraktionsprogramme von Xerox, Trados und U. Mügge, Oktober 2001. Last checked 28.9.2005 URL:
http://www.admin.ch/ch/i/bk/termdat/d/tworking/extract/extract.pdf

Streiter, O.; Zielinski, D.; Ties, I. and Voltmer, L. (2003): Term extraction for Ladin: An Example-based Approach. TALN 2003, Batz-sur-Mer, 11-14 juin 2003.

Thurmair, G. (2003): Making Term Extraction Tools Usable. Comprendium Germany. Last checked: 11.07.03. URL: http://www.comprendium.info/pic/papers/EAMT-2003-TExt-

article.pdf.

Warburton, K. (2001): LISA Terminology Survey Report 2001. Last checked: 28.06.05 URL: http://www.lisa.org/2001/termsurveyresults.html

Zielinski, D. (2002): Computergestützte Termextraktion aus technischen Texten (Italienisch), Saarbrücken: Saarland University. [Diploma thesis]. Last checked: 28.06.05. URL: http://fr46.uni-saarland.de/index.php?id=433

---

[1] URL of the survey: http://fr46.uni-saarland.de/t-survey/

[2] For the analysis of the results it has to be noted that not all participants answered all the questions nor did all participants complete the survey. The drop-out rate was around 15% (14% for translators and interpreters and 20% for terminologists and project managers).

[3] One project that currently addresses these issues is the eCoLoTrain project (Developing Innovative **eCo**ntent **Lo**calisation Opportunities for **Train**ers and Teachers in Professional Translation), a project funded by the Leonardo program of the European Union. For more information see http://fr46.uni-saarland.de/index.php?id=663.