Terminology extraction: an analysis of linguistic and statistical approaches

Maria Teresa Pazienza[1], Marco Pennacchiotti[1], and Fabio Massimo Zanzotto[2 1]

[1] Artificial Intelligence Research Group,
University of Roma Tor Vergata, Italy
{pazienza,pennacchiotti}@info.uniroma2.it
[2] University of Milano Bicocca, Italy,
zanzotto@disco.unimib.it

Abstract. Are linguistic properties and behaviors important to recognize terms? Are statistical measures effective to extract terms? Is it possible to capture a sort of *termhood* with computation linguistic techniques? Or maybe, terms are too much sensitive to exogenous and pragmatic factors that cannot be confined in computational linguistic? All these questions are still open. This study tries to contribute in the search of an answer, with the belief that it can be found only through a careful experimental analysis of real case studies and a study of their correlation with theoretical insights.

1. Introduction

The studies on the definition and implementation of methodologies for extracting terms from texts assumed since the beginning a central role in the organization and harmonization of the knowledge enclosed in domain corpora, through the use of specific dictionaries and glossaries [34]. Recently, the development of robust computational Natural Language Processing (NLP) approaches to terminology extraction, able to support and speed up the extraction process, lead to an increasing interest in using terminology also to build knowledge bases systems by considering information enclosed in textual documents. In fact, both Ontology Learning and Semantic Web technologies often rely on domain knowledge automatically extracted from corpus through the use of tools able to recognize important concepts, and relations among them, in form of terms and terms relations.

While terminology extraction (hereafter intended as the study of NLP based methodologies to extract terms from textual domain corpora) has found widespread application in Artificial Intelligence systems, the notion itself of *term* is still not clear, both from a pure linguistic and a computational point of view. Operatively it is thus possible to give only a general definition of term, as "a surface representation of a specific domain concept" [24],[30]. The difficulty in finding a deeper definition of term, in defining the properties that characterize univocally a term, and in

[1] This research has been developed during his sojourn at the AI Research Group at Roma Tor Vergata University

"translating" operatively these properties in a running system, still plays a central role in researches on computational linguistics. Recently, properties to define terms as *termhood* and *unithood* have been proposed in the literature [27], together with statistical measures and linguistic techniques able to "translate" such properties in computational algorithms.
In this study we present a few, commonly agreed, statistical and linguistic approaches used in NLP to extract and recognize terms, trying to compare their strengths in an automatic development environment. Moreover, we define a hybrid linguistic-statistical strategy that seems us to guarantee the extraction of a reliable terminology. Our approach has been implemented in a Terminology Extraction architecture (see later in Sec.3), whose aim is both to verify the validity of the many statistical measures proposed in literature and to evaluate existing and new linguistic methods for term recognition; as test bed a *spacecraft design* corpus provided by the European Space Agency has been used. As the issue on statistical measures adopted for recognizing terms is still matter of debate, in this study we will focus also on both methodological aspects and follow-up of the measures, trying to relate experimental evidences to their statistical methodological perspectives.
In Sec.2 we present and classify the different statistical, linguistic and hybrid approaches proposed in the literature, together with associated statistical measures and linguistic filters. In Sec.3 we describe our Terminology Extraction methodology, carefully comparing measures and linguistic techniques on a common test bed. Finally, in Sec.4 we try to outline conclusions and open questions.

2. Terminology extraction approaches

Current and past researches on computational terminology deal with a variety of approaches and strategies to extract and recognise terms by using both supervised and unsupervised techniques. Aim of most researches has been to obtain from a domain corpus the most significant set of terms, that is, the set of superficial representations of domain concepts that better represents the domain for a human expert.
In order to better understand and organize the work produced in the field, it can be useful to identify two mainstream approaches to the problem. From one side, statistical measures have been proposed to define the degree of *termhood* of candidate terms, i.e., to find appropriate measures that can help in selecting good terms from a list of candidates. From the other side, computational terminologists have tried to define, identify and recognise terms looking at pure linguistic properties, using linguistic filtering techniques aiming to identify specific syntactic term patterns. Finally, hybrid approaches try to use these two views together, taking into account both linguistic and statistical hints to recognise terms.
In this section we will present those that we regard as the main approaches adopted in the two mainstream view. Even if historically statistical approaches have been introduced before the linguistic ones, we firstly present the latter, since modern hybrid systems are usually composed by a cascade of a first linguistic analysis followed by statistical filters.

2.1 Linguistic approaches

Linguistic approaches to term recognition basically try to identify terms capturing their syntactic properties: in facts, it has been proved (see [8]) that terms usually have characteristic syntactic structures, called *synaptic compositions*: since the beginning, candidate terms have been mostly identified with noun phrases (e.g., the PHRASE system [17]).
Among researches that rely solely on linguistic analysis, in [9] it is postulated that syntactic data are sufficient to carry out term recognition. In this study the linguistic analysis is divided in two phases. Firstly, candidate terms are extracted using frontier markers that discard text sequences unlike to contain terms (such as phrases containing verb and pronouns). Then, relying on the analysis produced by a shallow syntactic parser, *parsing rules* are applied to the fragments survived to the first phase to select actual terms. Rules are created in an empirical way looking at experimental data. An example rule (for French) extracts from fragments of the type *[noun1 adj prep det noun2 prep noun3]* terms like *[noun1 adj noun2 prep noun3]*.
More recent works see the linguistic analysis simply as a set of *linguistic filters*, through which a system is able to retain admissible forms. Among other [3],[4] describe an approach to term extraction based on linguistic knowledge; moreover in [19] basic forms of English terms are *[noun, noun]* and *[adjective, noun]*, from which more complex syntactic patterns can be derived. In many works (such as [13],[26]) a simple *regular expression*s is supposed to be sufficient to identify the candidate terms forms.
In this direction a great effort has been done by [14]: in their extended study they try to identify the most common syntactic structures that terms assume, as inferred from the analysis of human produced terminological data banks. The study confirms the widely acknowledge intuition that terms generally appear in form of short noun phrases, mainly composed by only two *main items,* that is only two meaningful words, such as noun, adjectives (*adj*) and adverbs. These core terms, consisting of one or two main items, are called *base-terms*. The study identifies two major syntactic forms of base terms for English, *[adj noun], [noun noun]* , and three for French, *[adj noun], [noun noun], [noun prep noun]*. From the restricted set of base-terms more complex and long terms are formed via morphological or syntactic variations. Being thus the base-terms considered as forming the core terminology, most approaches in term recognition [13] focus only on them.
In such a view the extraction of candidate terms from a domain corpus is usually carried out as a cascade of two modules:

- A *parsing module,* able to perform a shallow linguistic analysis. Using Part of Speech (PoS) tagging techniques [10],[2] the module should guarantee the identification of *nouns*, *verbs*, *adjectives* and other part of speech in the text.
- A simple *term recogniser module*, that using regular expressions (or similar languages) extracts from the tagged text only the admissible surface forms, filtering out non interesting forms.

A debated issue on terminology recognition relates the identification of *term variations* [24]. As in [14], a term variant may be defined as *"an utterance which is*

semantically and conceptually related to an original term". For example the expression *lunar spacecraft mission* can be seen as a variant of the term *spacecraft mission*, conveying the meaning of the term augmented with another specific semantic information.
The study on term variants plays a role in term recognition, since particular type of variants can be seen as *transformed* forms of a term, that express exactly the same meaning of the related term (synonymy): for instance the variant *mission of spacecraft* is a "meaning-preserving" transformation of the term *spacecraft mission*. In case of meaning-preserving variations, in terminology recognition it can be justified to consider the original term and the variation as a unique term, "collapsing" the variation into the term. On the other side, non meaning-preserving variations can be seen as a way to identify complex terms built from base-terms. In [14] term variations are classified according to their characteristics. For term extraction, *permutations* (permuting a base-term with the *of* preposition, e.g. [*mission of spacecraft]*) assume a primary role, being one the strongest "meaning-preserving" transformation. Other interesting studies on the subject have been carried out by [24], devoted to the identification of particular kind of variants in perspective of a semantic structuring of terminologies.

Within a linguistic approach framework, other techniques can be applied in order to refine the terminology. For example a list of unwanted words (*stop-list*) can be used to discard those candidate terms that contain one of them. Usually the approach is to insert in the stop list function words and *generic words*, that is, words that are of very common usage in the language (for example "this", "that", "thing"). In most approaches stop-list words are automatically extracted from a generic corpus as those with the highest frequency, and are then validated by human experts. A stop-list can eliminate false terms consisting in generic collocations very common in the language, such as "*this thing"* or "*some day"* that being in the form *[adj noun]* could be likely selected as admissible surface forms.
To sum up, an ideal term recognition process within a linguistic approach should be able to:

- parse the domain corpus, identifying at least PoS;
- identify and extract candidate terms through admissible surface form rules;
- collapse meaning-preserving variations in the original term;
- implement other linguistic filters to refine the terminology.

What is produced at the end of the process is a list of good candidate terms likely to constitute the final terminology. However, a further analysis step is needed. In fact, the linguistic forms contained in the candidate terminology at this stage can be defined as filtered admissible surface forms, but not true terms. For example in a space domain candidate forms as *sufficient number* or *maximum size*, that are not specific domain expression, can easily survive the linguistic filters. What it needs is thus a step to select true terms from admissible surface forms. In other words, it must be implemented a sort of *termhood* definition in the process, able to discriminate among the surface forms. In pure linguistic approaches this process takes the form of a human expert manual validation. Unlikely, manual validation is not straightforward

as it seems (see Sec.2.4). The development of computational model able to capture the notion of *termhood* and to consequently identify true terms after the linguistic step, is thus clearly needed. Computational model usually consist in the application of statistical measures to the candidate term list, as described in the next session. The linguistic approach thus becomes a hybrid one.

2.2 Statistical approaches

Statistical measures applied to terminology are of great help in ranking extracted candidate terms according to a criterion able to distinguish among true and false terms and able to give higher emphasis to "better" terms. What is expected an ideal statistical measure could do is to assign higher scores to those candidates supposed to strongly possess a peculiar property characterizing terms. What is this *property* and what *"better"* means cannot be clearly stated: once again, an agreed definition of *termhood* could be helpful [34],[7].

Statistical approaches, like the linguistic ones, used alone only seldom reach truly satisfying results. While in pure linguistic approaches what lacks is a sort of "implementation" of *termhood*, the direct application of sole statistical measures to not-linguistically-filtered expressions can lead to a terminology rich of unwished forms. Indeed, only a few methods implement directly statistical measures without a syntactic-semantic analysis of the corpus. An example of pure statistical method is presented in [32], where 2-word candidate terms are extracted simply taking groups of two adjacent words, that are then weighted by the *Tf*Idf* statistical measure. In [25] sequences of words with length *N* are extracted, and then evaluated with an empirical measure based on term length and frequency.

In this section some of the major statistical measures for term recognition are described: our interest is in analyzing their effectiveness in combination with linguistic knowledge in hybrid approaches. Statistical measures can be classified by the following two distinct dimension: linguistic and statistical. A *linguistic dimension* is proposed in [27]: measures are divided in those that express *termhood* and those that express *unithood*:

- *Unithood:* expresses strength or stability of syntagmatic collocations.
- *Termhood:* expresses how much (the degree) a linguistic unit is related to domain-specific concepts.

By definition, *unithood* characterizes complex linguistic units (called *collocations*) composed by words with a strong association, such as compound words, idiomatic expression (e.g., *day after*) and complex terms (e.g., *spacecraft mission*). Therefore, *unithood*, while capturing an important aspect of terms, is not a peculiar property of them. Moreover being a measure of association, *unithood* is significant only for multiword terms, and cannot thus be applied to evaluate single word terms. On the contrary, *termhood* is a peculiar characteristic of terms, single word and complex.

The *statistical dimension* is based on statistic principles. Measures are classified accordingly to their methodological approach and the underlying assumptions[2] in:

[2] Classification proposed by the *Institut for Maschinelle Sprachverarbeitung , University of Suttgart* in www.collocation.de

Table 1. A classification of statistical measures in statistical and linguistic dimensions

		STATISTICAL DIMENSION		
		degree of association	***significance of association***	***heuristic***
LINGUIST. DIM.	***Unithood***	MI Dice Factor	z-score T-score X^2 Log Likelihood Ratio	MI^2 MI^3
	Termhood	"	"	Frequency C-Value Co-Occurrence

- *Degree of association* measures
- *Significance of association* measures
- *Heuristic* measures

A resuming classification graph is represented in Table 1, in which both dimensions have been depicted.
While *heuristic measures* are based on empirical and intuitive assumptions that often lack a theoretical statistical justification, the former two types of measures are usually based on a strong statistical background, as briefly described hereafter.
Association measures refer mainly to methods to estimate *unithood*. They are thus not used only in terminology, but in general for estimating collocations between two words[3] *u* and *v*, relying on the statistical evidence of occurrence of these words in the corpus. These evidences are expressed through a *contingency table* of *observed frequencies*, where *U* and *V* indicate respectively the first and the second words of the collocation. Co-occurrence of *(u,v)* is thus indicated by the frequency O_{11}, while *N* is the total number of collocation couples in the corpus ($N = O_{11} + O_{12} + O_{21} + O_{22}$).

	V=v	**V≠v**
U=u	O_{11}	O_{12}
U≠u	O_{21}	O_{22}

Moreover, *marginal frequencies* are defined as:

$R_1 = O_{11} + O_{12}$ $R_2 = O_{21} + O_{22}$ $C_1 = O_{11} + O_{21}$ $C_2 = O_{12} + O_{22}$

The aim of association measures is to draw inferences from the frequency table to estimate a collocation value. More in particular a *random sample model* is used, in order to generalize the observations in the frequency table of a single corpus (the *sample*) into assumptions valid for the language in general (the *population*). Consequently, being the measures an estimation, they will be prone to sampling errors. Specifically, what has to be estimated is a contingency table valid for the whole language, where X_{ij} are the frequencies of collocations in the whole language.

[3] All measures examined in the study are referred to the case of two-word terms, since they can be considered the most important and typical terms in a core terminology.

Assuming *independence* (occurrence of collocations are mutually independent) and *stationary* (the probability of seeing a particular word in the corpus does not vary) of the collocation event [16], values of X_{ij} can be derived from a *Bernoulli Distribution*, with τ_{ij} as *probability parameters* representing the probability that in the language the collocation X_{ij} outcomes in a single trial. It is then necessary to find an estimate of these values. Two ways can be followed: use a direct estimation of the parameters (as the *degree of association measures* do), or set some work hypotheses about them (as is the case of *significance of association measures*).
Degree of association measures estimate probability parameters from corpus evidences using *maximum-likelihood estimate* (MLE). Given the corpus frequencies, τ_{ij} are thus estimated as:

$$\tau_{11} \approx O_{11}/N \qquad \tau_{12} \approx O_{12}/N \qquad \tau_{21} \approx O_{21}/N \qquad \tau_{22} \approx O_{22}/N$$

Moreover, the probabilities of occurrence of the first and the second words in the language (respectively π_1 and π_2) can be estimated as:

$$\pi_1 \approx R_1/N \qquad \pi_2 \approx C_1/N$$

Combining the parameter estimates, different kind of measures can be derived. This approach is obviously prone to estimation errors, that are more likely to emerge when frequencies are low. To avoid estimation error caused by the MLE, *significance of association measures* try to calculate collocation using the *null hypothesis of independence, HI)*:

$$\tau_{11} \approx \pi_1 \cdot \pi_2$$

HI states that probability parameters π_1 and π_2 are independent. From the point of view of terminology thus means that there is not interesting relation between the two words composing the term. Under HI, using MLE of π_1 and π_2, it is thus possible to obtain the *expected frequencies* of collocation E_{11}, as the mean of the binominal distribution:

$$E_{11} = \tau_{11} \cdot N = \pi_1 \cdot \pi_2 \cdot N \approx \frac{R_1 \cdot C_1}{N}$$

HI is usually used by significance of association measures to compare the joint probability derived from a corpus with the joint probability in case of independence.

2.2.1 Statistical measure

In Table 2 the major statistical measures used in terminology recognition and evaluated in our study are presented.

Frequency doesn't derive from a theoretic statistical principle, but from the simple assumption that a frequent expression denotes an important concept for the domain in exam and should thus assume a high position in the rank of candidate terms. The most important objection in using frequency as a measure for term recognition concerns the fact that it doesn't take into consideration the degree of association (*unithood*) among

Table 2 Statistical measures and related formulae

MEASURE	ADOPTED FORMULA
Frequency	$f = O_{11}/N$
Church Mutual Information	$MI = \log_2(O_{11}/E_{11})$
Mutual Information variants	$MI^2 = \log_2(O_{11}^2/E_{11}) \quad MI^3 = \log_2(O_{11}^3/E_{11})$
Dice Factor	$DF = 2\frac{O_{11}}{R_1 + C_1}$
T-score	$TS = (O_{11} - E_{11})/\sqrt{O_{11}}$
Log Likelihood Ratio	$LLR = -2\log\frac{L(O_{11}, C_1, r)\cdot L(O_{12}, C_2, r)}{L(O_{11}, C_1, r_1)\cdot L(O_{12}, C_2, r_2)}$ where: $L(k,n,r) = r^k(1-r)^{n-k} \quad r = R_1/N \quad r_1 = O_{11}/C_1 \quad r_2 = O_{12}/C_2$
C-value	$CV = (len-1)\cdot\left(f - \frac{f(t)}{\lvert t\rvert}\right)$
Co-Occurrence	$CO = -\frac{\sum_N\sum_M O_{11i}}{\lvert N\rvert}$

words composing multiword terms [6]. Thus, very frequent expressions are considered good candidates while not being terms (e.g. *"this day"*). In order to capture indirectly the *unithood* nature of terms while using frequency, it is then necessary to implement linguistic filters able to discard candidates that don't have specific syntactic or morphological properties [26]. Frequency has been proved in several experimental studies (such as [13] and [28]) to be one of the most reliable measures for term recognition.

Mutual Information was originally defined in *information theory* [21], and then applied to linguistic analyses. In order to calculate Mutual Information it is necessary to estimate the probability parameters: in Table 2 we use the MLE, as proposed in [11]. A known problem of MI as presented in [10] is that it doesn't perform well with *low frequency* [16],[13]: in facts, the measure overestimates collocations composed by low frequency words. A solution to solve this problem, proposed in [11], is to exclude from the corpus collocations with frequency lower than a certain threshold. Another and more general solution is to find heuristic variants of the MI formula, such as $\mathbf{MI^2}$ and $\mathbf{MI^3}$ [13], that try to cope with low-frequency giving more importance to O_{11}, while laking a precise theoretic justification.

Dice Factor [33] suffers the low-frequency problem, as MI. In facts, DF is conceptually similar to MI, but, while the former theoretically derives from harmonic means, the latter is linked to geometric means.

T-score [12] rely on the *asymptotic hypothesis tests*, as other measures, such as Z-score [15]. The aim of T-score is to approximate the discrete binominal distribution (that is assumed to model collocations) with a distribution that converges to the continuous *normal distribution* for large *N*, relying on the null hypothesis of independence. Being a normal approximation of the binominal distribution, T-score suffers the well-known problems of *assumption of normality* [16].

Log-Likelihood Ratio [16] tries to solve the estimation problem of T-score and MI. The idea is to compare the probability of obtaining the contingency table observed in the corpus under the null hypothesis to the probability when there isn't independence, estimating the probability parameters τ_{11}, π_1 and π_2 with MLE and calculating the binominal distribution corresponding to the contingency table (*parametric test*).

C-value [22] is a linguistic based measure of termhood for multiword terms, that takes into consideration the frequency of the candidate term (*f*), the number of its main items (*len*) and information about how other candidates derived from the term are distributed in the corpus (*t* is the set of these candidates and *f(t)* their overall occurrences).

Co-Occurrence heuristically tries to capture *termhood*, relying on the assumption that a characteristic of terms is to co-occur in a same section of text with other terms (*N* are the corpus paragraphs in which the specific term appears, and O_{11i} the occurrences of the *M* terms in these paragraphs).

Other measures have been used for term recognition (but are not taken into consideration in our experiments): for example, *Tf*Idf* [23], *Domain Relevance & Domain Consensus* [6], *Contrastive measures* [34]. Many of these measure use a contrastive analysis of the domain corpus against a generic corpus (or many other specific corpora) in order to select terms.

2.3 Hybrid approaches

Recent terminology extraction systems combine linguistic and statistical techniques in structured hybrid approaches. Linguistic analysis is carried out before the application of statistical measures, to be helpful in selecting all linguistic admissible candidates over which will be applied numerical tests. Moreover, the reliability of a statistical measure increases when applied over linguistic justified candidates. The statistical step works on a list of candidate selected by the linguistic filters, trying to select and rank them according to a definition of *termhood* or *unithood* implemented through a specific measure.
One of the first systems using an hybrid approach is presented in [17], where noun phrases are firstly extracted as term candidates and then selected according to the frequency of their noun elements. In [13] linguistic candidates obtained by the application of syntactic patterns are filtered using different statistical measures, such as LLR, MI and frequency. In [26] a similar approach is followed: regular expression

are used in order to extract from the corpus linguistic candidates, that are then ranked by frequency.
A more complex architecture is envisioned in [18], where simple terms are firstly extracted according to frequency. New and more complex terms are then derived through linguistic heuristics and frequency filters applied to the simple terms retrieved in the first phase.
A step further is to deepen the linguistic analysis using semantic and contextual information. In [1] semantic information derived from thesauri, linguistic hints and statistical evidences are mixed together to rank candidate terms. For this purpose the *NC-value*, a complex heuristic measure, is proposed as a combination of C-value and of that *context-factor*, that takes into consideration the semantic, syntactic and statistical properties of the contexts in which the candidate terms appear.
The use of *extrinsic information* (e.g., contexts) is common also to other approaches. In [6] a shallow syntactic parser is used to select candidate term patterns; then Domain Relevance and Domain Consensus measures are applied to rank terms according to their contexts, intended at a wider domain level.
In [34] an *extensional definition* of term is proposed, in order to boost the term recognition process using frequency as a statistical measure, together with lexical and syntactic information about the contexts in which the term appears.

2.4 The evaluation issue

The evaluation of a term recognition system, as quality of extracted information, assumes a high relevance (further to performance evaluation) to both verify the validity of the underlying theoretical assumptions and to evaluate linguistic theories. Unlikely, even though automatic term extraction and recognition have a long tradition, no golden standards for evaluation have been introduced to clearly evaluate and compare different approaches.
The difficulty in outlining a generic and widely acceptable standard stems from the intrinsic nature of *term*. Indeed, as outlined in Sec.1, it is even difficult to give a precise linguistic definition of term. While an *operational* definition can be postulated, the problem remains for what concerns evaluation: then the need of a *golden standard* against which to measure systems performances. A golden standard can be provided directly or through validation only by a human expert. It is thus prone to the expert's subjective and personal interpretation of terms.
This layer of undetermination leads to more practical problems at a methodological level, where a method for evaluating an automatically extracted terminology is needed. Mainly two different methods are usually adopted for evaluation purposes: *reference list* and *validation*.
In the first case an a priori list of terms is assumed as a golden standard: in most cases the list is an already existing terminology for the specific domain. A reference list can also be constructed by a human expert examining the same corpus used for the automatic extraction. The quality performance of a system is evaluated in term of *Precision* (the percentage of extracted terms that are also in the reference list) and *Recall* (the percentage of terms in the reference list extracted by the system).

Validation method is preferred when a golden standard is not available or when particular characteristics of the extraction process have to be made explicit. In this case the performances are evaluated by a human expert that validates the terms extracted by the system. A *Precision* score is thus derived as the percentage of extracted candidate terms that have been retained as terms by the expert. Of course, manual validation is a time consuming activity. In [34] an account of what are the procedures and the difficulties in carrying out the process is given. In particular, manual validation requires two things. Firstly, the validation has to be done by more than one expert, in order to have the most reliable resource. Secondly, each expert must be introduced on the notion of what a term is: indeed, since the definition of *termhood* is pretty vague, it is likely that experts produce different validations, based on their own intuition of term.

Both methods have pro and cons. In terms of performance measures, the reference list technique is not the most suitable means to calculate *Precision*. In facts, it can happen that the system extracts true terminological expression that are not present in the reference list: while being good terms, these candidate are then recognized as false ones. From the other side, validation method is not able to capture *Recall*, since no other terms exist than those extracted by the system. Moreover, validation is a more system-dependent method, since it must be repeated for each system even when they operate on the same domain. Validation is also too much dependent on the personal judgement of the expert, that can be influenced in his validation task by external factor and by the list of terms already examined.

In the literature the problem of evaluation is still present, and maybe it will never be solved, thus limiting the development of an effective and standard framework in which to develop terms related technologies. In fact, since some systems adopt the reference list (e.g. [13]) and others the validation method (e.g. [20] [6]), it is impossible to clearly compare performances and thus to draw a precise line of evolution in term recognition methodologies.

3. Term recognition in practice: an hybrid approach

In order to override such an *empasse* we have carried on an in-deep analysis of the main methods used for term recognition in literature and cited in the previous sections. In particular we focus from one side on establishing a robust linguistic model to extract terminological expressions, and from the other side on evaluating and comparing different statistical measures when applied over. A wide debate is in fact active about the statistical validity and the mathematical foundation of many of the previously described measures (in particular those based on heuristic assumptions) [27]; a comparative study can be thus useful in order to understand their weaknesses, strengths, lacks and values. In such view, the overall term recognition process we envision can be classified as an *hybrid approach* composed by both a *linguistic* and a *statistical step*.

To evaluate different linguistic and statistical methodologies we tested our recognition process over a specific test bed. The corpus consists in a collection of domain specific documents related to *spacecraft design*, provided by the European

Space Agency (ESA) in the framework of the Shumi Project [31] jointly conducted by the AI Research Group of Roma Tor Vergata and the ESA/ESTEC-ACT (Advanced Concept Team). The collection comprehends 32 ESA reports, tutorials and glossaries, forming 4,2 MB of textual material (about 673.000 words). Once extracted, candidate terms have been validated by a team of ESA experts.

3.1 Linguistic step

As described in Sec.2.1 linguistic techniques to extract terms from textual corpora mainly consist on syntactic filters used to retain particular linguistic forms (i.e., syntactic patterns) as candidate terms. Moreover, stop-lists and term variations can be taken into consideration as further refinement. In order to examine these different techniques and to better understand the nature of terminology, we envision the linguistic step as an incremental process in which techniques performance are evaluated. Firstly, we extracted from the corpus those linguistic forms corresponding to specific syntactic patterns (*admissible surface forms*) (Table 3) considered as good prototypes of candidate terms, that are classified in *k-word* categories, where *k* indicates the number of *main items* contained in the term.

Table 3. Syntactic patterns used to extract *k-words* candidate terms represented in RegExp

Terms length	Syntactic patters
1-word	(noun)
2-word	(adj)(noun) (noun)(noun) (noun)(prep)(noun)
3,4,5-word	(noun){3,5} (noun)(prep)(noun){2,4} (adj)(noun){2,4}

In order to carry out the term extraction process we previously analyzed the corpus document using a modular syntactic parser [5] together with a dedicated term extraction module [31]. Out of the 44.619 candidate terms extracted, 6346 have been retained as true terms by the ESA experts, leading to an overall Precision of 14%. Considering only terms which appear in the corpus more than 5 times Precision increases to 38%, giving a first indication that frequency could be an interesting measure to select terms.
Then, all the 44.619 candidate terms have been filtered using a generic *stop-list* of *specific determiners* (definite articles, demonstrative and possessive adjectives) and *general determiners* (indefinite articles and expressions as *few, many, some*, etc.). The aim is to discard a priori candidates that, by definition, can not be considered terms. In facts, determiners are generally defined as "non-descriptive words that have little meaning apart from the nouns they refer to". As terms should be formed only by meaningful words, candidates containing determiners should be discarded. The stop-list[4] has been automatically derived as the most frequent determiners extracted from a

[4] The stop-list comprehend the following words: *this, all, some, these, such, any, many, both, those, each, same, own, another, few, several, least, every, more, fewer, much, there, most*.

generic human-annotated sub-corpus of the British National Corpus. After the *determiners stop-list* passage, 2556 candidates are filter out, increasing Precision from 14% to 15%, while Recall decreases only to 99,3%.
A second *adjective stop-list* composed by a list of validated most frequent 200 adjectives of the sub-corpus has been applied in order to verify the value of candidate terms containing generic common adjectives; the intuitive hypothesis is that common adjectives such as *same, another, industrial, next, available, military* are not enough significant to define a term. Results show a slight increase in Precision (18%) while Recall drops to 81%.
A complete list of results using stop-lists is reported in Table 4, both for all terms and for the subclass of 2-word terms. As it can be noticed the subclass of 2-word terms has an higher Precision, whose motivations will be discusses later on. In general, the use of stop-lists seems to improve Precision, having as side effect a decrease in coverage, mostly for terms of more than two words. In the rest of the study the set of terms obtained after the two stop-list filtering will be used for the analysis. It consists of 28.465 terms (among which 5134 validated as true terms) whose characteristics are summarized in
Table **5** (excluding 21 spurious terms).
A first conclusion from our study can be at this point already drawn: 2-word terms seems to be the most important and frequent terms (as already outlined in Sec. 2.2 and [14]), as out of the 5134 true terms 3150 (61,4%) are 2-word. This result is in line with previous analysis carried out in [14], where 56% of terms contained in a hand collected terminology bank are 2-word. For the scope of this study we will thus hereafter focus mainly on *2-word* terms retained after the stop-lists filtering.
An interesting analysis relates the syntactic structure (i.e. syntactic patterns) of the 2-word terms extracted and validated. In Table 6 the characteristics of 2-word terms classified by syntactic patterns are shown (of course, referring specifically to English, for other languages different values can be expected). The reported statistic takes into consideration *inflectional variations*, that is, singular and plural forms of nouns are collapsed to a unique term (e.g. *spacecraft(s) mission(s)*). The most common terms are those of the form *adjective-noun*, followed by forms *noun-noun*, both in the extracted set (column 2) and in the true (i.e., validated) terms set (column 4). Examples of frequent *noun-noun* terms are *application datum*, *test level, source packet*; frequent *adjective-noun* are *magnetic field, solar wind, technical requirement*. Fewer terms have the form *noun-prep-noun* (for instance *speed of light*, *factor of safety, satellite in orbit*), most of which have *"of"* as preposition. Our results are fairly in line with those obtained in [14].

Table 4. Precision and Recall using stop lists

	All candidate terms		2-word candidate terms		
	Precision	Recall	Precision	Recall	F-Measure
Before stop-lists	14,2%	100%	43,6%	100%	60,7%
After det stop-list	15%	99,3%	44,2%	100%	61%
After adj stop-list	18%	80,9%	47,1%	86,5%	61%

Table 5. Characteristics of terms as obtained after stop-list processing. Precision is intended as the number of correct terms (*column 4*) over the total number of terms of a certain class (*column 2*)

Term class	n. of terms	% over the total	n. of correct terms	% of correct over total correct	Precision
1-word	6625	23,3 %	1177	22,9 %	17,8 %
2-word	16369	57,6 %	3150	61,4%	19,2 %
3-word	4229	14,9 %	697	13,6 %	16,5 %
4-word	978	3,4 %	102	2 %	10,4 %
5-word	243	0,8 %	8	0,1 %	3,3 %

Table 6. Characteristics of validated 2-word terms by syntactic patterns. Precision is intended as *column 4* over *column 2*

Syntactic pattern	n. of terms	% over 2-word	n. of correct terms	% of correct over all 2-word correct	Precision
adj noun	7122	43,5 %	1363	43,3 %	19,1 %
noun noun	4714	28,8 %	1206	38,3 %	25,6 %
noun prep noun	4022	24,6 %	548	17,4 %	13,6 %
spurious	511	3,1 %	33	1 %	6,4 %

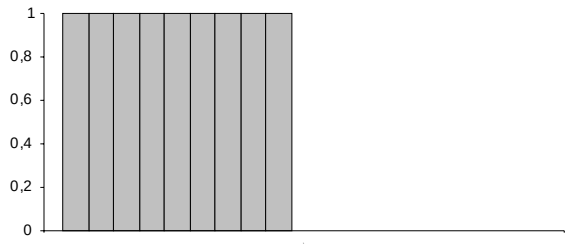
Even though our study has been applied to only one domain, it can be a first indication on the performance of linguistic approaches over different syntactic patterns for English. It is interesting to notice, for example, that *noun-noun* terms while constituting only the 28,8% of extracted 2-word terms, are the 38,3% of true 2-word terms, having the overall highest Precision (25,6%). That seems to point out that *noun-noun* forms are more promising than the others.
For what concerns the issue on *term variation* already discusses in Sec.2.1, we decided to leave the problem aside. In our view it is difficult to build an a priory methodology based on a linguistic theory able to justify the collapse of term variants in a base term, even in apparently obvious cases such the *"of" permutation*. Collapsing a variant assumes in fact that the variant and the base term convey the same meaning, that is not always true. For example in the *"of"* case we found variant-term couple such as *list of definition – definition list* and *field of view – view field* which are not completely meaning preserving. The only exceptions exists for *inflectional variants*, since singular-plural variations on nouns can be roughly considered meaning preserving. In the literature some term recognition approaches take into consideration variations (e.g. [13]) while other prefer to left the problem aside as we do [26].

2.2 Statistical step

In our approach the set of terms produced by the linguistic analysis is input to a successive statistical process, willing to rank terms according to their *termhood* or *unithood* properties. Our statistical analysis is twofold.

From one side, a wide debate is still going on what could be the most suitable measure for ranking and selecting terms. In fact, since it is impossible to define an objective and widely accepted golden standard/benchmark for measuring terminology, it is clearly difficult to establish measure performances and accuracy. As far as we know, only a few studies tried to compare the most adopted measures on a common test bed (e.g.,[13] [28]). In our view it is thus necessary to test the different measures on many different domain corpora in order to clarify and solve this issue.
From the other side, we aim to point out the different characteristics of the measures we tested, willing to identify which properties of terms they let emerge from the ranks they produce.
As test bed for the measures cross-evaluation we use the set of 2-word terms obtained after the linguistic analysis (terms extraction and stop-list filtering). In particular tests will be applied to the 949 terms with a frequency $f \geq 5$ (e.g., terms that appear in the corpus 5 or more times). As suggested also in [13] and [20] it is evident that the choice of using a frequency threshold over 2-word candidates seems to be the best compromise to obtain a functional set of terms for evaluating measures over a clean test bed. In facts, as demonstrated in [28], statistical methods perform badly when applied to very low frequency objects.
As golden standard for evaluation we use the set of true terms validated by the ESA experts among the total of 949. True terms are 447, leading to an overall Precision (after the linguistic step) of 47,1%.
We evaluate measures in two steps. Firstly, we apply the method used in [13]. Terms are ranked according to a specific measure and then divided in equivalence classes of 50 consecutive elements in the ranking. For each class Precision is calculated as the percentage of correct terms in the class. In this view the best statistical measure should be the one able to clearly separate true terms from false ones: that is, the *ideal* measure should assign the highest positions in the rank to the 447 true terms, leaving the remaining false terms to the lowest part of the ranking (see Fig.1). In such a way what is evaluated is *the power of each measure in discriminating true and false terms*.
As a second evaluation we simply use the standard method of plotting Precision of a given measure at different Recall percentiles. Here, Recall is defined as the percentage of true terms contained in a ranking interval over the total 447 true terms. Precision is thus the percentage of true terms at a given Recall percentile over the total number of terms at the same percentile.

Fig.1. The curve of an *ideal* measure to rank terms, with Precision of the measure in the *y* axis at different equivalence classes (*x* axis). Equivalence classes are order by increasing value of the measure

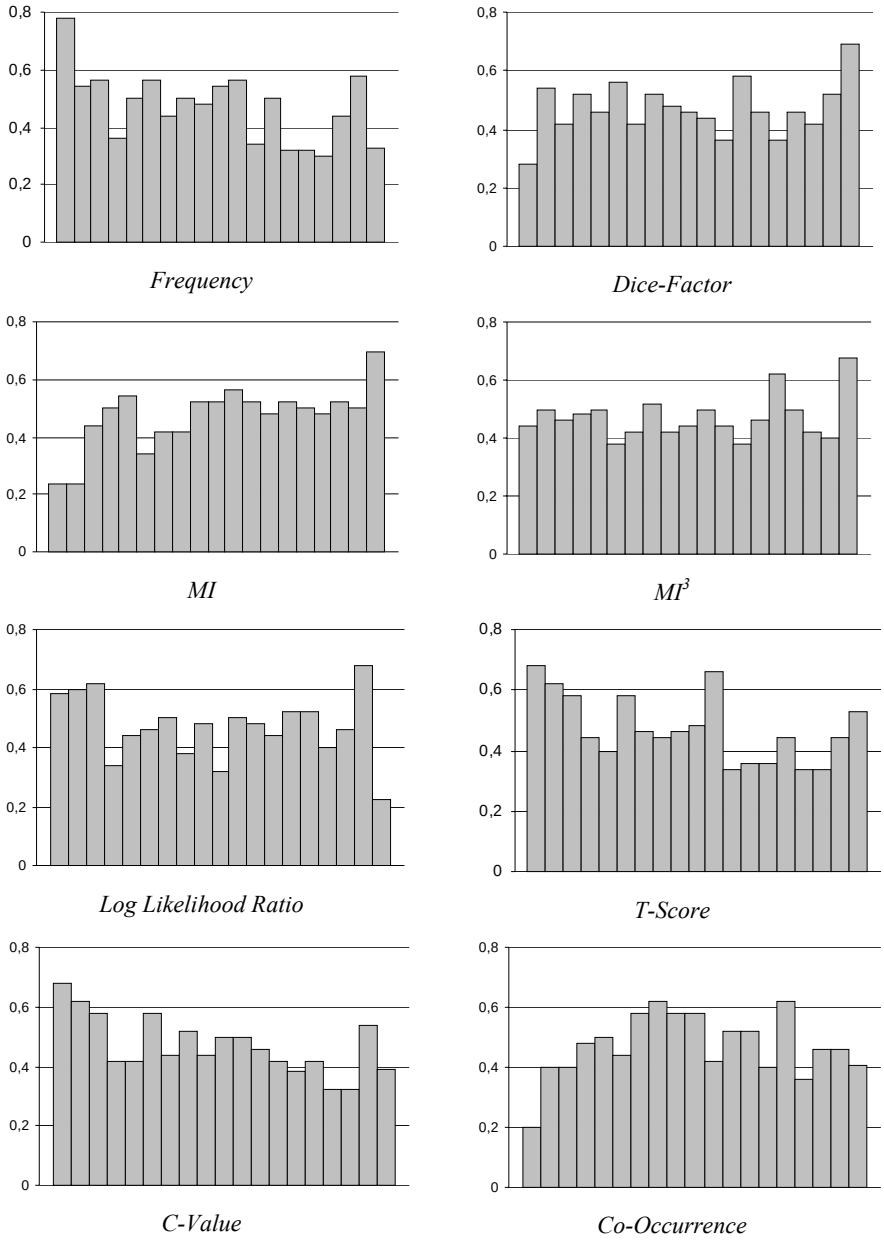


We compared some of the most widely used measures for term recognition, focusing on those that need only information about the specific domain in order to be calculated. That is, we don't take into consideration measures such as Tf*Idf or Domain Relevance that need some sort of corpora comparison. In fact, while comparing the *lexical profiles* of the relevant domain against a generic domain (or a set of different domains) appears to be useful in term recognition (since the definition of *term* itself underlines the importance of domain specificity), we want here to restrict our attention to the simplest (and more likely) cases in which only a domain corpus is available. The compared measure are thus: *frequency, T-score, MI, MI^3 , Dice Factor, Log Likelihood Ratio (LLR), C-value* and *Co-occurrence*.
Results are summarized in the histograms in Fig.2, in Fig. 3 and Table 7.

2.2.1 Analysis of results

By a first look to histograms in Fig.2 it emerges that a pool of measures seems to have an interesting behaviour, compared to what should be the ideal measure. In particular, *frequency*, *C-value* and *T-score* have an overall decreasing trend, indicating that for lower values of the measure Precision decreases. This behaviour suggests that these three measures tend to assign higher value to true terms: so, the better a term is ranked by the measure, higher is the probability of being a true term. *Frequency*, *C-value* and *T-score* seem thus measures able to discriminate in some way among terms and to produce a significant rank.
Other measures show approximately a flat curve, thus revealing to be poor statistics for recognizing terms. In particular it is interesting to notice how *degree of association measures* (i.e., *MI* and *Dice Factor*) are characterized by a curve that grows in the first equivalence classes. That is, these measures tend to behave badly in the higher part of the rank (many of the terms with highest score are false terms). The reason of this behaviour lies in the already mentioned (see Sec.2.2) problem of low frequency that afflicts *MI* and *Dice Factor*: these measures give a too high score to rare events (e.g. to terms composed by rare words) (that could be useful for recognizing very rare terms appearing in large document collections).
To clarify Table 8 shows the first 20 ranked terms by *MI* and *frequency*, together with the occurrences values of the term words in the corpus. As it can be noticed, *MI* tends to rank higher terms composed by words with low frequency: those terms, while having a high association score, are usually not interesting, since they are very rare linguistic expressions of the corpus. On the contrary, the most relevant terms according to *frequency* have been successfully validated by the experts, suggesting that a recurrent expression is in fact a good term.
Comparing the histograms of *MI* and MI^3 it can be noticed how the latter measure seems to act successfully in removing the problem of low frequency (as indicated in [13]) for the first equivalence classes; notwithstanding, MI^3 doesn't seem to be interesting anyway, being characterized for the rest by the same flat curve as MI.
Interestingly, *LLR,* that has been proved to be a useful measure for term recognition in other studies (e.g.,[13],[16]), doesn't seem to give the same indication in our experiment (see not well characterized curve in Fig.2) even though it presents a slightly decreasing trend.

Fig.2. Precision of different measures (*y* axis) at different equivalence classes (*x* axis). Equivalence classes are order by increasing value of the measure



Summing up, there isn't a measure that presents an histogram function comparable to the ideal curve, but however, some of them are able to produce a rank in which the probability of having correct terms is higher in higher position of the rank.

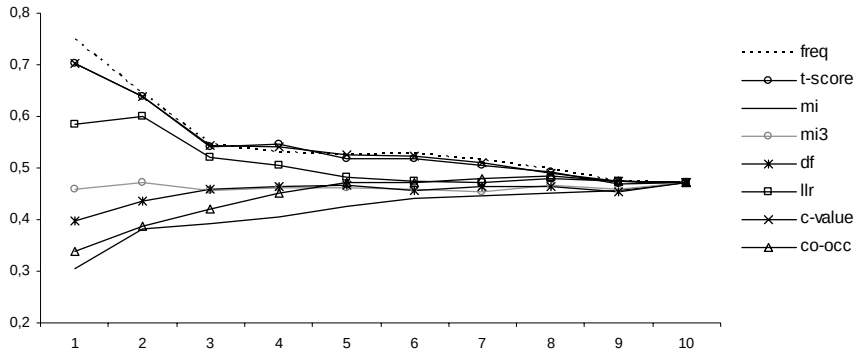


Fig. 3. Overall Precision of different measures (*y* axis) at different Recall percentiles (*x* axis)

Table 7. Precision at different Recall percentiles for statistical measures. In grey the best value at each specific percentile

RECALL PERC.	MEASURES							
	freq	*t-score*	*mi*	*mi3*	*df*	*llr*	*c-value*	*co-occ*
0,1	75,0%	70,3%	30,4%	45,9%	39,8%	58,4%	70,3%	33,8%
0,2	64,3%	63,8%	38,1%	47,1%	43,5%	60,0%	63,8%	38,8%
0,3	54,7%	54,2%	39,4%	45,8%	45,9%	51,9%	54,4%	41,9%
0,4	53,1%	54,5%	40,4%	46,0%	46,5%	50,6%	54,1%	45,2%
0,5	52,4%	51,8%	42,6%	46,1%	46,6%	48,3%	52,4%	47,1%
0,6	52,7%	51,7%	44,2%	46,0%	45,6%	47,4%	52,2%	47,3%
0,7	51,5%	50,6%	44,7%	45,3%	46,5%	47,2%	51,0%	47,9%
0,8	49,7%	49,3%	45,2%	46,6%	46,4%	47,9%	49,0%	48,4%
0,9	47,4%	46,9%	45,7%	46,0%	45,5%	47,5%	47,3%	47,5%
1	47,1%	47,1%	47,1%	47,1%	47,1%	47,1%	47,1%	47,1%

Results obtained for the second evaluation are reported in Fig. 3 and Table 7. A first analysis of the Precision curve reveals a neat distinction in two curve classes. Indeed, a group of measures starts with a high Precision (between 60-75% at the first percentile) and then decreases quite substantially. On the contrary, a second group starts with very low Precision (between 30-45%) and then slightly increases. It is interesting to notice that the first group comprehends *frequency, T-score, LLR* and *C-Value*, the second *MI, MI*^{3} , *Dice Factor* and *Co-occurrence*. The first group is thus composed by measures strongly based on frequency (*C-value* and *frequency* itself) and *significance of association measure* (*T-score* and *LLR*). All these measures outperform the second group almost at all percentile, indicating that frequency and the statistical null-hypothesis of independence assumption are better means to rank and recognize terms compared to the probability parameters approximation methods used by the *degree of association measures*. The neat low values of some of these latter measures at the first percentiles is again an evidence of their low-frequency problem.

Table 8. 20 higher ranked 2-word terms by *MI* (*left*) and by *frequency* (right). *R1, C1* and *O11* are respectively the occurrences of the first word, the second word and the term. Experts validation (*True* or *False* terms) is in the last column.

TERM	O_{11}	R_{11}	C_{11}	val
tape_recorder	6	6	1	F
extension_of_maximum	6	9	3	F
additive_for_processing	6	8	3	F
scan_platform	6	9	4	F
circuit_board	6	7	4	T
adaptive_routing	7	10	4	T
capacity_of_spur	5	12	7	F
nic_fluctuation	5	14	9	F
audible_noise	6	12	7	F
million_of_dollar	5	6	7	F
industry_association	12	13	3	F
cleaning_agent	5	10	8	F
destination_identifiers	5	9	8	F
remote_sensing	5	5	12	T
statement_of_effectivity	9	18	9	F
accordance_with_subclause	15	19	4	F
imaginary_circle	5	12	10	F
pound_of_payload	8	12	9	F
behavioural_view	5	14	11	F
look-up_table	6	20	14	F

TERM	O_{11}	R_{11}	C_{11}	val
application_datum	122	581	510	T
magnetic_field	104	246	231	T
solar_wind	101	119	483	T
technical_requirement	83	1000	1098	T
test_level	69	355	677	T
source_packets	61	147	173	T
source_datum	60	581	609	F
normative_document	59	108	53	F
technical_specification	58	156	304	F
launch_vehicle	53	104	140	T
mechanical_part	50	142	187	T
mission_phase	50	135	267	T
test_requirement	48	1000	1365	T
performance_requirement	47	1000	1014	T
user_manual	46	65	28	F
flight_operation	43	351	387	T
propulsion_system	42	402	386	T
gray_system	41	402	500	F
sub-service_provider	40	43	22	F
engineering_process	39	204	264	T

Also from this second evaluation method *frequency* emerges as the best measure, since its Precision is higher at almost all percentiles, while the worst measure appears to be *MI*. Moreover, it emerges that theoretically similar measures such as *MI, MI^3* and *Dice Factor* have different behaviours. In particular, MI^3 performs better at the beginning (thanks to the solved low-frequency problem) and then becomes similar to *Dice Factor*, while *MI* seems to remain quite apart at lower Precision values.

Considering the results of the two evaluations it can be noticed how *significance of association measures* (*T-score* and *LLR*) perform better than *degree of association measure* (*MI*, MI^3 and *Dice Factor*), while a few of the *heuristic measures* have good performances (such as *frequency*). In theory, it could be justified by the different statistical methodologies that *degree* and *significance measures* use to calculate the association score: it would thus emerge that it is better to adopt methods that use the null hypothesis of independence rather than those that try to only approximate probability parameters with MLE.

For what concerns the other statistical dimension, no final conclusion can be drawn about the statistical behaviours of measures of *termhood* and *unithood*, since measures curves don't seem to be characterized by these properties. Notwithstanding, an interesting linguistic analysis is to compare the highest ranked terms by the best measures of *termhood* and *unithood*, in order to see how look like terms with high *termhood* and terms with high *unithood*.
In Table 9 the first 20 terms are reported for the best measure of *termhood* (*frequency*) and the two best measures of *unithood* (*LLR* and *T-score*). At first glance it can be noticed how the first three terms in the rank are common for the three measures, while

Table 9. 20 higher ranked 2-word terms by *frequency*, *LLR* and *T-score*

FREQ	LLR	T-score
application_datum	magnetic_field	application_datum
magnetic_field	application_datum	magnetic_field
solar_wind	solar_wind	solar_wind
technical_requirement	normative_document	technical_requirement
test_level	abbreviated_term	test_level
source_packets	user_manual	source_packets
source_datum	sub-service_provider	source_datum
normative_document	source_packets	normative_document
technical_specification	launch_vehicle	functional_test
launch_vehicle	electromagnetic_radiation	technical_specification
mechanical_part	architectural_design	abbreviated_term
mission_phase	mechanical_part	launch_vehicle
test_requirement	technical_specification	electromagnetic_radiation
performance_requirement	parameter_statistic	mechanical_part
user_manual	telecommand_packets	mission_phase
flight_operation	mission_phase	test_requirement
propulsion_system	logical_address	performance_requirement
gray_system	minimum_capability	user_manual
sub-service_provider	pressure_vessel	flight_operation
engineering_process	functional_test	propulsion_system

going down in the ranking, agreement decreases, suggesting a certain stability among measures in selecting higher terms. *Frequency* has 17 terms in common with *T-scores*, and only 11 with *LLR*, while T-score and *LLR* 13: that seems to confirm no practical importance in the measures classification into the linguistic dimension *termhood-unithood*.

In conclusion, *frequency* appears the best measure (as confirmed in [13] and [20]), followed by *T-score* and *C-value*. *LLR* doesn't show good performances as in other studies, while behaving better than *MI*, MI^3 and *Dice Factor,* whose recognition power seems substantially poor. *Co-occurrence* poor results appear to indicate that at a first analysis information about terms co-occurrence in text is not an interesting property to distinguish true form false terms; that is, terms don't seem to have the property of appearing together, concentrating in specific section of texts.

Taking into consideration computational complexity, the position of *frequency* gets even stronger, being its computational cost irrelevant, since frequency can be calculated as the occurrence of terms in the corpus during the linguistic step. The other interesting measures, *T-score C-value* and *LLR*, while being comparable to *frequency* in term of recognition performance, are not from a computational point of view. Their good recognition power is thus overridden by their computational cost.

4. Conclusions

In this paper it has been widely analyzed the problem of term recognized task in an automatic process, also by considering the continuously growing interest in terminology as a useful hint for ontology learning as well as for supporting Semantic Web. This required to converge to an operational definition of *term* (to be effective in

an extraction system) and to agree on the need of both linguistic and numerical knowledge for systems with such an ability.
The minimal set of needed linguistic process has been underlined and described in a general architecture for terminology extraction. Then , a large set of widely adopted statistical measures have been applied and comparatively evaluated in order to determine their role in improving terminology extraction. A real corpus has been used to produce a list of candidate terms and a related evaluation carried on them has been possible thanks to a parallel manual evaluation produced by human experts.
The overall system performances have been compared with state of the art results, showing its higher reliability. As a last point, authors will underline that with our approach we are able to recognize (in the processed corpus) linguistic expressions that are real terms while not being validated by the expert interested in a tight specific application domain (e.g., *tape recorder*). This is due to the fact that assuming *"the corpus as containing only terms related to the application domain"* is not totally correct: the jargon of the writers covers, in facts, a wider context than the specific domain of interest.

References

1. Ananiadou S., Maynard D.: Identifying contextual information for term extraction. In Proc. of 5th International Congress on Terminology and Knowledge Engineering (1999)
2. Basili, R., Pazienza M.T.,Velardi P. :An Empirical Symbolic Approach to Natural Language Processing. Artificial Intelligence, vol. 85 (1996)
3. Basili R., De Rossi G., Pazienza M.T.:Inducing Terminology for Lexical Acquisition. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2),Brown University, Providence, Rhode Island (1997)
4. Basili R., Bordoni L., Pazienza M.T.: Extracting terminology from corpora, Proc. of the 2nd International Conference on Terminology, Standardization and Technology Transfer (1997)
5. Basili, R., Pazienza, M.T., Zanzotto, F.M.: Customizable modular lexicalized parsing. In: Proc. of the 6th International Workshop on Parsing Technology (2000)
6. Basili R., Missikoff M., Velardi P.: Identification of relevant terms to support the construction of Domain Ontologies, ACL workshop on HLT, Toulouse, France. (2001)
7. R. Basili, M. T. Pazienza, F. M. Zanzotto: Decision trees as explicit domain term definition 19th International Conference on Computational Linguistic (COLING2002). Taipei (Taiwan) (2002)
8. Benveniste, E.: Problèmes de linguistique générale. Gallimard (1966)
9. Bourigault, D.: Surface grammatical analysis for the extraction of terminological noun phrases. In: Proc. of Fifteenth International Conference on Computational Linguistics (1992)
10. Brill E.: Some advances in transformation-based part-of-speech tagging. In Proceedings of the 15th International Conference on Computational Linguistic, 1034-1038 (1994)
11. Church K.W., Hanks P.: Word Association Norms, Mutual Information and Lexicography. ACL 1989, 76-83
12. Church, K. W., Gale E., Hanks P., Hindle D.: Using statistics in lexical analysis. In Lexical Acquisition: Using On-line Resources to Build a Lexicon, Lawrence Erlbaum. (1991)
13. Daille, B. : Approach mixte pour l'extraction de termilogie: statistique lexicale et filters linguistiques. PhD Thesis, C2V, TALANA, Universitè Paris VII (1994)
14. Daille, B., Habert, B., Jacquemin, C., Royaut, J. :Empirical observation of term variations and principles for their description. Terminology, 3(2) (1996) 197-258

15. Dennis, Sally F: The construction of a thesaurus automatically from a sample of text. In Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation, Washington, DC. (1965) 61-148
16. Dunning T.: Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19(1) (1994) 61-74
17. Earl, L.L. : Experiments in Automatic Extracting and Indexing. Information Storage and Retrieval 6(X) (1970) 273-288
18. Enguehard C., Pantera L.: Automatic Natural Language acquisition of a terminology. Journal of Quantitative Linguistics 2(1) (1994) 27-32
19. Evans D.A., Zhai C.: Noun-phrase analysis in unrestricted text for information retrieval, Proceedings of the 34th conference on Association for Computational Linguistics. Santa Cruz, California (1996) 17-24
20. Evert S., Krenn B.: Methods for the qualitative evaluation of lexical association measures. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France. (2001) 188-195
21. Fano R. M.: Transmission of Information: A statistical Theory of Communications. MIT Press, Cambridge, MA. (1961)
22. Frantzi K.T., Ananiadou S.: Extracting Nested Collocations. COLING 1996. 41-46
23. Hisamitsu T., Tsujii J.: Measuring Term Representativeness. Third Summer Convention on Information Extraction (SCIE 2002). Roma, Italy (2002)
24. Jacquemin, C.: Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Universitè de Nantes, France (1997)
25. Jones L.P., Gassie E.W., Radhakrishnan S.: INDEX: The statistical basis for an automatic conceptual phrase-indexing system. Journal of the American Society for Information Science 41(2) (1990) 87-97
26. Justeson, J., Katz S.: Technical Terminology: some linguistic properties and an algorithm for identification in text. In: Natural Language Engineering, 1 (1995) 9-27
27. Kageura K., Umino B.: Methods of automatic term recognition. Terminology, 3(2). (1996)
28. Krenn B.: Empirical Implications on Lexical Association Measures. Proceedings of The Ninth EURALEX International Congress. Stuttgart, Germany. (2000)
29. Nakagawa H., Mori T.: Automatic term recognition based on statistics of compound nouns and their components. Terminology 9(2):201 (2003)
30. Pazienza, M.T.: A domain specific terminology extraction system. In: International Journal of Terminology. Benjamin Ed., Vol.5.2 (1999) 183-201
31. Pazienza, M.T., Pennacchiotti, M., Vindigni, M., Zanzotto, F.M.: Shumi, Support To Human Machine Interaction. Technical Report. ESA-ESTEC contract N.18149/04/NL/MV – Natural Language Techniques in Support of Spacecraft Design (2004)
32. Salton, G., Yang, C.S., Yu, C.T.: A Theory of term importance in automatic text analysis. In: Journal of the American Society for Information Science 26(1) (1975) 33-44
33. Smadja F.A., McKeown K., Hatzivassiloglou V.: Translating collocations for bilingual lexicons: a statistical approach. Computational Linguistics, 22:1. (1996)
34. Zanzotto, F.M.: L'estrazione della terminologia come strumento per la modellazione di domini conoscitivi. PhD Thesis, Università degli Studi di Roma Tor Vergata (2002)