IST-2000-25338 PEKING
Deliverable Identification Sheet

Project ref. no.	*IST-2000-25338*
Project acronym	**PEKING**
Project full title	**People and Knowledge Cross Lingual Information Gathering**

Security (distribution level)	*Public*
Document name	*D3.1. Database filling devices and documentation*
Type	*Deliverable*
Status & version	*"Final version"*
Number of pages	*23*
WP contributing to the deliverable	*WP3*
WP / Task responsible	*gilcUB*
Other contributors	*KUN, CRF, Quinary, Meta4*
Author(s)	*N. Bel, M. Villegas, M. Marimon - gilcUB* *K. Koster - KUN* *L. Gilardoni - Quinary*

DOCUMENT EVOLUTION (optional)

Version	Date	Status	Notes
0.1	15/02/2002	Draft	
0.2	25/03/2002	Draft	

0.3	1-04-2002	final	

TABLE OF CONTENTS

0. EXECUTIVE SUMMARY

The role of WP3 in PEKING Project is to provide PEKING functionalities such as document classification, KAT extraction and document retrieval with linguistic methods to improve accuracy and to allow handling documents in different languages. In PEKING 'cross-linguality' refers to the different linguistic processes required to work with documents and texts written in different languages.The present document is a summary of the work carried out on the tasks T3.1 to T3.4 within the work package 3 in the PEKING Project. WP3 tasks T1 to T4 were concerned with the development of a lexical database containing different types of expressions (Nouns, Noun Phrases and Verb Phrases). Monolingual expressions were to be extracted from already existing lexical and terminological resources, as well as a result of extracting processes from the texts to be used for classification. These extracted units will be used for creating a multilingual lexical resource.

Within PEKING several tools are to be used for Linguistic Processing. This is due to the fact that these tools have been developed for different languages outside the project. Thus, we will refer to the general processes that these tools have to carry and we will include as appendixes the descriptions of the actual systems used for each language.

For the purpose of building the multilingual database to be used within PEKING, linguistic processing tools apply to convert a text into different units that will represent it. These units will be Linguistically Motivated Terms.

This document is concerned with the acquisition of LMT's from the texts PEKING is to deal with and its storage for later cross-lingual linking: In section 2, a first definition of the lexical database will be given. In section 3 a motivation for the use of LMT's will be supplied as to justify the need of linguistic processing. Sections 4 and 5 report about the work done for the monolingual components, and section 6 will report about the need for using defined and new standard formats for ensuring usability of different results for different components.

1. Introduction

This document is a summary of the work carried out on the tasks T3.1 to T3.4 within the work package 3 in the PEKING Project. WP3 tasks T1 to T4 were concerned with the development of a lexical database containing different types of expressions (Nouns, Noun Phrases and Verb Phrases). Monolingual expressions were to be extracted from already existing lexical and terminological resources, as well as a result of extracting processes from the texts to be used for classification.

The role of WP3 in PEKING Project is to provide PEKING functionalities such as document classification, KAT extraction and document retrieval with linguistic methods to improve accuracy and to allow handling documents in different languages. In PEKING 'cross-linguality' refers to the different functionalities to work with documents and texts written in different languages.For **document classification** it is widely acknowledged that information embodied in single words is not sufficient for statistical methods to perform accurately. Monolingual linguistic methods are to supply them with units different than single words to improve its accuracy. These units are to be called 'Linguistically Motivated Terms', and they can be multiword expressions (collocations, compounds), or abstractions over linguistic data (ordered pairs of items). Besides, 'Linguistically Motivated Terms', for different languages will be clustered under the translation equivalent relation in a Multilingual Database. Thus, classification and KAT extraction will be able to work with different languages.

As for **cross-lingual document retrieval**, cross-lingual search is mostly based on the translation of terms for different languages. For identifying those terms some linguistic processing must be done. In PEKING, a pull event will involve a monolingual query string which needs to be translated into a multilingual boolean search expression. This process implies identifying sufficient and necessary 'key-words', assigning them a LMT and a translation equivalent for a given domain, sense, etc. The Multilingual Database will support the translation process.

Within PEKING several tools are to be used for Linguistic Processing. This is due to the fact that these tools have been developed for different languages outside PEKING, and as they require long development processes, it was outside the scope of this project to attempt a unified view of such tools. Thus, we will refer to the general processes that these tools have to carry and we will include as appendixes the descriptions of the actual systems used for each language.

The goal of Linguistic Processing is to take a text and deliver a representation of it that can be handled by other PEKING components: Text Classifier, KAT extraction and Document Search and Retrieval. To that end the tasks to be done can be summarized as follows:

Segmentation: Identification of the units a text is made of.

Tokenization: Identification of the nature of the units found in a text, i.e. words, punctuation marks, figures, etc.

Lexical look-up: access to a lexical repository in order to find out the linguistic information associated to each word.

Lemmatization and disambiguation: identification of the linguistic characteristics of a word in a given context.

Chunking: Identification of the different component that form a multiword unit by linguistic or statistical means.

Term identification: Identification for a given domain and in a given text of words or multiwords units whose referent is more restricted than for the same word or multiword units in a general context.

Parsing: Identification of the syntactic relations that hold among different words, i.e. phrases, sentences, subject, object, etc.

Transduction: transforming the result of chunking or parsing into a particular form, Head/Modifier frames, suitable for further processing:

> Head/Modifier frame (HM frame, HMF): pair of the form [head, modifier], possibly nested, i.e. containing further (embedded) HMF's.
>
> Unnesting: transforming a nested HMF into a set of atomic HMF's.
>
> Atomic HM frame: HM frame containing no embedded HMF's.

For the purpose of building the multilingual database to be used within PEKING, these different processes apply to convert a text into different units that will represent it. These units will be ***Linguistically Motivated Terms*** (LMT).

This document is concerned with the acquisition of LMT's from the texts PEKING is to deal with and its storage for later cross-lingual linking: In section 2, a first definition of the lexical database will be given. In section 3 a motivation for the use of LMT's will be supplied as to justify the need of linguistic processing. Sections 4 and 5 report about the work done for the monolingual components, and section 6 will report about the need for using already defined standards for ensuring the re-use of different results by different components.

2. Database characteristics

For the multilingual database, we will be using a dB generator developed by gilcUB. The dB generator creates a relational dB and a dB interface out off a relational model expressed in a SGML DTD. It also includes procedures to automatically load dtd-conformant data encoded in SGML into the dB. For PEKING DTD see Appendix 1.

3. Items to fill the database: Linguistically Motivated Terms for cross-lingual classification and content search

The selection of terms to represent the contents of texts or its *aboutness* has been one of the most important areas of research in the field of Information Retrieval [Bruza and Huibers 1996]. The rationale behind is to find those indexing terms that should describe the concepts present in the documents.

As for Text Classification using Machine Learning methods, the observed characteristics of a set of documents classified under a given class are used to induce an 'automatic classifier' which will decide for a new document whether it has the characteristics required in order to be classified under that particular class. Text categorization uses basically the same methods than Information Retrieval (IR), and thus shares the problem of how to represent texts. Most systems take a document as a vector of term (also called *features*) weights that represent how much a given term contributes to the contents of a document. For most of the systems, these terms are the words the document is made of.

In IR has been reported that individual words do not characterize the contents of text and several attempts have tried with different ways of dealing with more sophisticated representations for texts. Some systems experimented with *phrases* rather than single words as indexing terms [Fuhr et al. 1991; Schütze et al 1995; Tzeras and Harmann 1993], being phrases units linguistically or statistically motivated. Results reported have not been conclusive and the predominant feeling [Spark Jones, 1999] is that only 'shallow' linguistic techniques like the use of stop lists and lemmatization are of use in IR. Simple linguistically motivated techniques turned out to be no more effective than well-executed statistical approaches, while more advanced NLP techniques, such as concept extraction are too expensive for large-scale IR applications.

Some other systems take advantage of pre-existing domain specific thesauri and glossaries to identify pertinent terms, which can be made of more than a word, [Betts, 1991] showing better quality for characterizing the contents of a text or its 'aboutness'. However the existence of such thesauri is still scarce and the manual maintenance and upgrade of terminological databanks is still a marginal and expensive exercise which cannot cope with the growing number of texts in different domains. It was clear that existing glossaries and terminological resources for human use does not cover all the requirements for classification (neither for cross-lingual search, as we will see later). Thus a new line of interest in developing methods for automatic term extraction arose joining efforts coming

from different fields. Devices for (semi-)automatic acquisition of terms have been addressed using linguistic and quantitative methods in the last decade.

Most of these technical terms to be identified are multiword noun phrases and differ from other noun phrases because their meanings "are not unambiguously derivable from the meanings of the words that compose them" [Justeson and Katz, 1995]. Following Justeson and Katz, several linguistic and quantitative evidences allow to identify technical terms. These evidences have been used to build hybrid methods (linguistic and statistic) to automatically identify them. For a survey of current systems see [Cabré et al. 2001].

Following this line, shallow linguistic processing techniques as well as introducing NP phrases identification have been tried for improving Text Classification systems. However, the results of the different experiments made [Apté et al. 1994; Dumais et al. 1998; Lewis 1992] in the past have not shown clear better results as far as classification purposes are concerned. [Caropreso et al. 2001] is one of the last experiments for text categorization using phrases understood as statistically relevant sequences of words or n-grams. Their work is a study of the usefulness of statistical phrases for TC independently of the learning algorithm used. In assessing their results we must take into account that, as required by most TC methods, they had to do feature or term selection.

Most document classification techniques are crucially based in automatic feature/term selection. The main difficulty these techniques face is the high dimensionality of the space: terms are words or phrases that occur in documents. Depending on the collection, terms can be hundreds of thousands. Most of these techniques rely on quantitative methods for selecting those terms which will be used for classification computations. [Caropreso et al. 2001] concluded that although the use of bigrams did not lead to better results, one central issue that remains open is the relation that feature selection could have had in the obtained results. Other related aspect raised by [Caropreso et al. 2001] as crucial to asses different classifiers is how found bigrams, for their experiments, have to affect the frequency of its parts. Their suspicion is that important unigrams are pushed out of the selected features list by bigrams that incorporate them.

As reported in [Yang & Pedersen 1997] most commonly used feature selection techniques deal with the following key issues:

- favoring common terms or rare terms
- task-sensitive or task-free: includes category information
- using term absence to predict the category

[Yang & Pedersen 1997] report from their experiments, that the TC methods that performed better are those that favored the most common terms in their final selected lists. However, one might argue that the phenomena behind is that low frequency terms cannot help in classifying as they show up very little in documents. That is, phrases used as terms are more precise than single words as they benefit from the mutual disambiguation effect of words, but the probability for a specific phrase to re-occur in different documents is smaller than that of words. This conclusion is in line with [Lewis 1992], who argued that changing

the representation of texts affects their statistical qualities: there will be lower document frequency for terms. These facts have also been attested by various authors working in IR as [Arampatzis et al. 2000] that note that the use of phrases may greatly improve precision, but at a drastic price in recall.

Thus, linguistically motivated terms means that 'linguistic' units, phrases, can be useful for classifying as they help in capturing the 'aboutness' of documents, i.e. 'resources' is not to be considered the same in the case of 'human resources' than in the case of 'financial resources' or 'material resources'. However, by doing this we are changing the quantitative characteristics of texts, and this changes must be taken into account for term selection and later use in TC. By 'chunking' different words into a n-gram we will be affecting frequency of words in documents and the dimensional space:

Frequency counting of 'resources' in ILO corpus[1]

term-A	*term-B*	*#-A*	*#-B*	*#A+B*
human	resources	1970	801	224
financial	resources	1000	801	209
material	resources	338	801	37

Space: from 4 terms to 7 terms

resources + human + financial + material +
human resources + financial resources + material resources

As noted by [Caropreso et al. 2001] two central issues are:

- to go further in assessing to what extent identifying a phrase as 'human resources' is of any help if it is not finally selected because its lower frequency and the impact that its creation has in the dimensional space.
- to evaluate whether the chunking of two terms in a NP-term has to impact the frequency counting of the former two terms, i.e. for the example above final counting will be

term	*#before*	*#after*
human	1970	1566
financial	1000	791
material	338	301
resources	801	331
human resources	-	224
material resources	-	209
financial resources	-	37

[1] Despite of the different methods used for feature selection, given that those based in frequency are allways among the three with better results with little differences, for the sake of simplicity we shall be taking only frequency measures for exemplification.

Taking into account what has been said, linguistic processing is to be used to support term selection techniques. Two novel techniques are under investigation in PEKING:

1) the systematic use of morphological, syntactic and semantic normalization to conflate terms that are both linguistically and statistically related (fuzzy matching) [Koster et al., 1999]

2) the development of feature selection techniques suitable for HM phrases. Traditional stop-list techniques do not apply. A novel Term Selection technique, based on Uncertainty, has been elaborated and has been evaluated in comparison with other local TS techniques [Peters and Koster, 2002].

KUN is presently investigating the statistics of HM frames, in the context of the SBC algorithm, which is mathematically well tractable. Thus, for improving text classification in PEKING, we will be using Linguistically Motivated Terms to assist the feature selection process.

PEKING is also studying the benefit of using LMT's for multilingual classification. The hypothesis behind is that selected LMT's from documents pertaining to the same area but in different languages will result in comparable sets of terms. The terms extracted from documents in one language can be mapped to the ones found in another language. Thus documents to be classified will become a list of LMT's suitable for the same classifier.

Cross-lingual search in PEKING also requires the linking of terms that are "translational equivalents" for a given domain. Thus, the multilingual database could be accessed by both functionalities in order to get translations. We are to describe briefly the characteristics of the terms used for cross-lingual search and to what extent the multilingual database can be shared by both components.

Mapping of terms in different languages can be done at a first instance by accessing terminological repositories and bilingual glossaries already available (see 4.2.5). However we know from cross-lingual information retrieval applications that translation lexicons available are very limited. In the CLIR domain, used translation lexicons are made of bilingual dictionaries but in order to enlarge its coverage some applications have tried to extract translation equivalences from aligned corpora. As mentioned by [Resnik, Oard and Levow 2001], both approaches show some drawbacks. While bilingual dictionaries and glossaries provide reliable information, they propose more than one translation per term and don't include translation preference information. Aligned corpora for very innovative domains, as technical ones, offer contextualized translations, but the errors introduced by statistical processing of texts in order to align them are considerable.

Another area where text alignment has been done massively for translation purposes is Translation Memory systems. Translation Memories are translation support tools that maintain a database source and a target language sentence pairs, and automatically retrieves the translation of those sentences in a new text which occur in the database. Recent experiments with that kind of tools, as the one made by [Simard and Langlais, 2001], have shown that most used algorithms for text alignment (based on number of characters) are counterproductive for aligning units shorter than sentences. On the other hand it seems that using segments with a clear syntactic status, such as phrases, gives better results as these are likely to have a clear identifiable translation which will show more occurrences. Thus, these authors use statistical means to decide what is the most probable translation based on frequency.

For PEKING we will be testing this approach. For that purpose we will select Linguistically Motivated Terms which are nominal phrases for cross-lingual search. [Jutseson & Katz 1995] collected evidences that technical terms are mostly Noun-Noun compounds in English, although also quantitatively relevant samples of Adjective-Noun, Adjective-Adjective-Noun, Adjective-Noun-Noun, Noun-Adjective-Noun, Noun-Noun-Noun, and marginally Noun-*of*-Noun deemed consideration. As for Spanish is commonly acknowledged that most terminological multiwords units are made of Noun-prep[2]-Noun, Noun-Adjective. As in English, less frequent, but nevertheless important, are phrases made of Noun-Adjective-Adjective, Noun-Adjective-prep-Noun. This decision will allow the 'semi-automatic' mapping of phrases in one language to another overcoming the problem of limited coverage of existing resources.

The multilingual database is being filled first with exiting bilingual resources, as said. These resources already offer the translation of Noun Phrase terms. A second step will be to introduce NP LMT's as supplied by the linguistic processing tools. These will only be introduced if a translational equivalent is found in the list of LMT's for the other language. In order to find such a translational equivalent a compositional translation based on all possible translations of the words that form the phrase will be done, taking into account as basis lemmatized words for lexical look-up. As said before this gives rise to many possible translations. Only the phrase (or phrases) that is found to be a LMT (in the list of selected ones) will be introduced. For example:

fijación de salarios

fijación => fixing, securing, fastening, posting, establishing
salarios => wage, pay, salary

Possible compositional translations:

wage fixing, wage securing, wage fastening, wage posting, wage establishing
pay fixing, pay securing, pay fastening, pay posting, pay establishing
salary fixing, salary securing, salary fastening, salary posting, salary establishing

[2] the most frequent prepositions is *de*, but also *con* and *por* are found.

Only occurrence in English LMT's *wage fixing* that appears 372 times in the English corpus.

More experimentation with that approach is required to test the performance that will be done in the next period.

4. Already existing resources: glossaries, computational lexicons, terminological repositories and corpus selection

4.1. Corpus collection

Objective of this task was to collect texts for different purposes:

1) wordform identification for monolingual lexicon completion
2) identifying representative texts within the proposed domains for grammar extension's phenomena definition
3) experimentation for Monolingual Classification and KAT extraction and for Multilingual Classification

Users were asked to supply with texts that will be used for classification experimentation purposes. In addition a multilingual, classified corpus was to be found in order to carry multilingual experiments.

4.1.1. Spanish

CINDOC put a collection of classified texts at the disposal of the project. Corpus analysis (lemmatization and corpus statistics) was done.

CINDOC CORPUS: 16.388 documents

However most of these files only contained text in capital letters and without accents, making difficult its use for linguistic processing. A subset of this corpus, i.e. those files that had normal text was used.

Total number of suitable files: 4584

The total number of words for non-full capitalized files is 612413, with an average of 133 tokens per document.

4.1.2. Dutch

CORPUS FutD

The corpus from FutD consists of about 31000 documents of legal/fiscal prose, pre-classified into 19 main classes and 1118 subclasses (a hierarchical multi-classification). The corpus has a size of 730 Mbytes (106 M words, about 3000 words per document avg).

4.1.3. Italian

CRF Corpus

It is made of 58 classified documents which amount a total of 25086 words. Documents are technical reports which show specific characteristics of technical documents, as the presence of English terms in Italian texts.

4.1.4. Multilingual corpus

CRF Bilingual Corpus

It is made of 101 classified texts. 58 are Italian and 43 English. Italian texts amount a total of 25086 words and English texts 19878.

Documents are technical reports being from 400 to 1700 words length.

ILO CORPUS:

It comes from a database which is classified according to the following 12 classes:

02 - Human rights
03 - Conditions of employment
04 - Conditions of work
05 - Economic and social development
06 - Employment
07 - Industrial relations
08 - Labour Administration
09 - Social Security
10 - Training
11- Special provisions by category of persons
12 - Special provisions by Sector of Economic Activity

The document's content is regulations, conventions, general studies, comments by experts from the International Labour Organization. We had a look to them and seem reasonable in terms of length and variety. The actual database concerning languages is made of parallel documents (translations) but we have mixed the docs, that is, we have included some which can be found both in English and Spanish, and some that aren't.

Document distribution per class: ranging from 80 docs to 200 Document length: minimum 117 words, maximum 7.500 words. Most of documents are among 2000 and 3000 words.

4.2. Monolingual Lexicons

Linguistic processing modules used require linguistic information about words in texts. Although such lexicons existed for the languages deal with in the projects[3], the coverage of existing resources had to be tuned to the current applications both in what refers words coming from user domains, and in information, adding linguistic information not present such as framing. To that end, some work had to be done for completing monolingual lexicons for the different languages so as to allow accurate linguistic preprocessing of all documents as well as the implementation of robustness techniques for ensuring information based on linguistic guessing or on probabilities.

From the corpus collected for each user, a wordform list was extracted for each language. This list was checked against existing monolingual lexicons as to extract missing words, or, for those words already present in monolingual lexicons, checking that information is complete according to grammar requirements, and correct for the area or domain of end users[4].

Besides, existing glossaries and terminological repositories were collected.

4.2.1. Dutch

A large lexicon of Dutch (250 000+ entries), based on CELEX and other sources, providing morphosyntactic and syntactic information, as well as word stem. Containing detailed subcategorization information for verbs, and some for adjectives, and the beginning of a collection of idiomatic stock phrases.

The AMAZON grammar of Dutch, developed by Van Bakel et al. from the Linguistics department, has been adapted to Information Retrieval applications, including verb categorization in order to be able to determine the functions of the sentence constituents, and the transduction to HM frames. The grammar includes robustness devices at the phrasal and lexical level, and therefore has a high coverage and no special domain lexica is required for the FISCAAL application.

4.2.2. English

A large lexicon of English (200 000+ entries and 78 000 collocations), based on WordNet 1.6 and various other sources, providing morphosyntactic and syntactic information but no lemmata. Subcategorization information for verbs and adjectives and lexical frequencies obtained from tagged corpora have been added.

---

[4] It could be the case that some words are present but in aceptions / senses that do not correspond to the use of this particular word in a particular domain. For instance: *break* can be in a database as a verb but not as a noun, which will be the right part of speech in a sentence as: We will have a break at 12:00.

The EP4IR (English Phrases for IR) grammar was developed by KUN for Information Retrieval purposes. It includes subcategorization and transduction, and contains robustness devices at the phrasal and lexical level, so that in parsing a stream of (not necessarily syntactically correct) text the most probable HM frames are obtained and the use of lexical probabilities is presently being implemented.

4.2.3. Italian

FACILE tools lexicon has shown to have an acceptable coverage of the texts delivered by CRF. Lexical gaps have been estimated between 5 and 7% of the total items [Ferraro & Gilardoni 2002].

As an addition resource, PEKING is counting with ItalWordnet, a large semantic database (~80K synsets) developed within an italian national project, SI-TAL, aimed at realizing a set of integrated resources and tools for the automatic processing of the Italian language. Within SI-TAL, ItalWordNet is the reference lexical resource which will contain information related to about 130,000 word senses grouped into synsets. This lexical database is not being created ex novo, but extending and revising the Italian lexical wordnet built in the framework of the EuroWordNet project. The lexical coverage of ItalWordnet is being extended by adding a terminological subset belonging to the economic and financial domain.

CRF delivered the company's Monolingual Glossary with about 800 entries with definitions.

4.2.4. Spanish

gilcUB lexicon contains 65,000 lemmas with morphological and morphosyntactic information. Framing for disambiguation purposes has been added for verbs (about 6.000 lemmas). Due to the low rate of unknown words (less than a 10% including capitalizations without accents, misspellings and foreign terms), robustness methods have been included guess possible categories of different items have been integrated. This was required in order disambiguation rules to apply properly.

4.2.5. Bi-lingual glossaries

A total of 8.000 items, English-Spanish translation correspondences list has been collected for the economical domain to be used within PEKING.

ITAL-Wordnet (see 3.2.3) gives the possibility to link Italian and English. Italian synsets are linked to English concepts through a cross lingual index (ILI), although more research in this respect is required.

5. Linguistic Processing tools for acquiring terms

Linguistic tools to be used in PEKING are:

- AGFL for English and Dutch, KUN
- TAGIT and RTAG for Spanish, gilcUB
- FACILE tools and RTAG for Italian, Quinary

5.1.1. KUN linguistic processing tools

The AGFL system is a parser generator system for natural languages, which has been developed at KUN with support from the Dutch Research Organization NWO and the NLnet Foundation. It is presently being brought under the GNU Public Licence. Parsers generated by AGFL are subject to the Lesser GPL, which makes it possible to embed them as a component into academic and commercial software, without licensing costs. More about AGFL at its website: http://www.cs.kun.nl/agfl/.

The linguistic department of KUN, which also uses AGFL, has provided tagger/lemmatizers for English and Dutch, which are used for deriving lexical frequencies. Scripts for the construction and maintenance of lexica are available, and will be brought into the public domain along with the AGFL system. Also see Appendix 2 for a short description of

5.1.2. gilcUB linguistic processing tools

TAG-IT: Text tokenizer for Spanish: it carries out delimitation of basic textual units, identification and classification of elements, identification of special constructions (dates, numbers, etc.), lexical lookup.

RTAG is a lexical disambiguation tool based on analysis promotion/reduction by means of weighted symbolic context rules. It also selects non-recursive chunks according to defined patterns in a regular expression grammar. See Appendix 3 for tool documentation.

5.1.3. Quinary linguistic processing tools

FACILE processing tools perform the following processes for PEKING [Gilardoni et al. 2001]:

The **tokeniser** identifies the start and end position of the token and the characters in it (token).

The **morphological analyser** adds a set of morphological analyses for each token and its normalised form (norm).

The **tagger** selects a single part-of-speech tag for each token (syn), and by comparing this with the morphological analyses, it is possible to filter out those that are not compatible with the selected tag.

6. Representation Issues: Standards

Initiatives such as those of EAGLES/ISLE; MATE, SALT in the definition of standards for the encoding of lexical resources have been of great help and with no doubt further taken into account for the PEKING developments. Thus the consortium have taken profit of the already developed resources from the different groups already following such guidelines. PEKING proposal has thus taken advantage of 'de facto' standards which have been used by the members such as EUROWORDNET and national ITALWORDNET, a compatible model with the already mentioned EAGLES/ISLE standards. As the development of lingware is now linked to the idea of sharing resources instead of rebuilding it for each application, PEKING is taking measures for ensuring that standards are used for new description of items to be worked with in the project.

7. References

Apté, C., F.J. Damerau and S.M. Weiss (1994), Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12.

A. Arampatzis, Th.P. van der Weide, C.H.A. Koster, P. van Bommel (2000), Linguistically-motivated Information Retrieval. Encyclopedia of Library and Information Science, volume 69, pages 201-222, Marcel Dekker, Inc. - New York - Basel, 2000.

Betts, R. and D. Marrable (1991), Free Text vs. controlled vocabulary, retrieval precision and recall over large databases. In *Online Inf 91*. London.

Bruza, P. and T.W.C. Huibers (1996), A Study of Aboutness in Information Retrieval. *Artificial Intelligence Review*, 10.

Cabré, M.T., R. Estopà and J. Vivaldi (2001), Automatic Term Detection: A review of current systems, in D. Bourigault, Ch. Jacquemin and M-C. L'Homme (2001) *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam.

Caropreso, M.F., S. Matwin, F. Sebastiani (2001), Statistical Phrases in Automated Text Categorization.

Dumais, S.T., J. Platt, D. Heckerman and M. Sahami (1998), Inductive learning algorithms and representations for text categorization. In *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management*.

Ferraro, M. and L. Gilardone (2002), Linguistic coverage test on CRF corpus, PELING Internal Working Document, March 2002.

Gilardoni, L., C. Biasuzzi, M. Ferraro (2001), Tools Preliminary Survey, PEKING Technical Document.

Koster, C.H.A, (July 2001) Frames and the unnesting process. WP1 - PEKING working document.

Koster, C.H.A., C. Derksen, D. van de Ende and J. Potjer (1999), Normalization and matching in the DORO system. *Proceedings BCS-IRSG 1999 colloquium*, Glasgow University.

Lewis, D.D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of SIGIR-92, ACM International Conference on Research and Development in Information Retrieval*.

Peters,C. and C.H.A. Koster (2002), Uncertainty-based Noise Reduction and Term selection in Text Categorization, *Proceedings ECIR 2002*, Glasgow March 25-27, Springer LNCS 2291

Resnik, P., D. Oard and G. Levow (2001), Improved Cross-Language Retrieval using Backoff Translation. In *Proceedings of HLT2001*, San Diego, California.

Simard, M. and P. Langlais (2001), Sub-sentential Exploitation of Translation Memories. In *Proceedings of Machine Translation Summit* VIII, Santiago de Compostela, Spain.

Yang, Y. and J.O. Pedersen (1997), A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning.*

8. APPENDIXES

8.1. APPENDIX 1 - PEKING Multilingual Database DTD

```
<!DOCTYPE PEKING [

<!ELEMENT PEKING - O ( Monolingual+ , Bilingual+)>


<!ELEMENT Monolingual - O ( MorfUnit? , Feat?, SemUnit?)>


<!ATTLIST Monolingual
        lexiconname         CDATA #REQUIRED
        language            CDATA #REQUIRED
        version             CDATA #IMPLIED
        creationdate        CDATA #IMPLIED
        modificationdate    CDATA #IMPLIED
        property            CDATA #IMPLIED
        copyright           CDATA #IMPLIED>

<!ELEMENT MorfUnit - O ( MorfDesc?, HasMeaning?)>

<!ATTLIST MorfUnit
        id              ID              #REQUIRED
        spelling        CDATA           #IMPLIED
        gramcat         (NOUN|ADJ|VERB|LOC)      #REQUIRED
      gramsubcat   (MAIN|AUX|COMMON|PROPER) #REQUIRED>


<!ELEMENT MorfDesc - O EMPTY>
<!ATTLIST MorfDesc
      morfunit    IDREF #REQUIRED
      feat        IDREF #REQUIRED>

<!ELEMENT HasMeaning - O EMPTY>
<!ATTLIST HasMeaning
      morfunit    IDREF #REQUIRED
      semunit           IDREF #REQUIRED>


<!ELEMENT Feat - O EMPTY>
<!ATTLIST Feat
        id        ID                #REQUIRED
        feature         CDATA                   #REQUIRED
        value           CDATA                   #REQUIRED
        comment         CDATA                   #IMPLIED >
```

```
<!ELEMENT SemUnit - O ( SemDesc?)>

<!ATTLIST SemUnit
        id              ID              #REQUIRED
        spelling        CDATA           #IMPLIED
        comment         CDATA           #IMPLIED>

<!ELEMENT SemDesc - O EMPTY>
<!ATTLIST SemDesc
      semunit     IDREF #REQUIRED
      feat        IDREF #REQUIRED>


<!ELEMENT Bilingual - O ( BilingualUnit? )>


<!ATTLIST Bilingual
        lexiconname         CDATA #REQUIRED
        language            CDATA #REQUIRED
        version             CDATA #IMPLIED
        creationdate        CDATA #IMPLIED
        modificationdate    CDATA #IMPLIED
        property            CDATA #IMPLIED
        copyright           CDATA #IMPLIED>


<!ELEMENT BilingualUnit - O EMPTY>
<!ATTLIST BilingualUnit
        id            ID #REQUIRED
        sourceunit          IDREF #REQUIRED
        targetunit          IDREF #REQUIRED
        comment             CDATA #IMPLIED>

]>
```

8.2. APPENDIX 2 KUN Linguistic Processing tools-

[please insert here document D3-1annex2.pdf]

8.3. APPENDIX 3 gilcUB Linguistic Processing Tools

Bel, N., A. Chillarón, M. Marimón and J. Porta. RTAG-4.2. gilcUB. March 2002.

[Please, insert here document D3-1annex2.pdf]

8.4. APPENDIX 4 - Quinary Linguistic Processing Tools

Fabio Ciravegna, Alberto Lavelli, Nadia Mana, Luca Gilardoni, Silvia Mazza, Massimo Ferraro, Johannes Matiasek, William Black, Fabio Rinaldi, David Mowatt. "Classifying Texts Integrating Pattern Matching and Information Extraction", In *Proceedings of the 16th Inter. Joint Conference on Artificial Intelligence* (IJCAI99), Stockholm, 1999.

[please insert here document D3-1annex4.pdf]