Applying Light Natural Language Processing to Ad-Hoc Cross Language Information Retrieval

Christina Lioma, Craig Macdonald, Ben He,
Vassilis Plachouras, and Iadh Ounis

University of Glasgow, G12 8QQ, UK
{xristina, craigm, ben, vassilis, ounis}@dcs.gla.ac.uk

Abstract. In the CLEF 2005 Ad-Hoc Track we addressed the problem of retrieving information in morphologically rich languages, by experimenting with language-specific morphosyntactic processing and light Natural Language Processing (NLP). The diversity of the languages processed, namely Bulgarian, French, Italian, English, and Greek, allowed us to measure the effect of system-specific features upon the retrieval of these languages, and to juxtapose that effect to the role of language resources in Cross Language Information Retrieval (CLIR) in general.

1 Introduction

The driving force behind our participation in CLEF 2005 has been to explore the effect of morphologically rich languages across a set Information Retrieval (IR) platform, in terms of system-specific features and language resources. From the outset it was anticipated that this effect would be considerable, not only from a computational perspective, i.e. technical implementation issues involving character encodings, but most importantly with reference to the availability and quality of language resources provided for the said languages, such as stemmers and lexica.

This year's language selection formed a representative sample of some of the major branches of the IndoEuropean family of languages, spanning from the Slavonic branch, to the Latin, the Germanic, and even the Hellenic branch. We used the same retrieval platform as reported in CLEF 2004 [6], on top of which we added selective language-specific Natural Language Processing (NLP).

This paper is organised as follows. Section 2 presents an overview of the linguistic foundations of this work, with special note being made to the language processing approaches adopted. Section 3 presents and discusses our monolingual and bilingual runs. Section 4 concludes with a summary of the approaches tested and the extent of their success.

2 Linguistic Background

Natural Language Processing is considered essential to the retrieval of highly inflectional languages, of rich morphology and syntax. The validity of this statement has been tested for Greek-English IR. Moreover, noun phrase extraction,

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 170–178, 2006.

a popular NLP application that purports to capture constituent structure, and thus add an extra dimension to the conceptual content of a given text as rendered by single words, has been put to the test for monolingual French and bilingual Italian-French retrieval. Noun phrase identification and extraction has been realised using our in-house Noun Phrase (NP) extractor, which identifies noun phrases on the basis of their syntactic features alone, and independently of corpus statistics. The said NP extractor, which can currently process English, French, Italian, and German, is designed to target both nested and discontinuous noun phrases, with an adjustable maximum thershold of terms allowed between members of a broken noun phrase, and no limitations with regards to the length of a given noun phrase.

Additional NLP applications utilised in the context of this work include light syntactic analysis, achieved by a probabilistic part-of-speech (POS) tagger, lemmatisation, and morphological analysis [10,14]. Unfortunately, the unavailability of such technology for Bulgarian meant that only French, Italian, English, and Greek were subjected to this type of examination. The part-of-speech tagsets used for the aforementioned languages adhere to the Penn TreeBank Tagset conventions [7], with a few exceptions. These exceptions stem from the fact that languages are not always syntactically isomorphic. The collective part-of-speech tags used are presented in Table 1. The initials in square brackets relate to the specific language to which the said tags are exclusive. DE, EN, FR, GR and IT stand for German, English, French, Greek and Italian respectively. The class distinction refers to the linguistic distinction between function words and content words [4]. Tags falling under the closed class are assigned to words bearing very little to nil content. Such words are peripheral to the semantic load of their environment, and exist mainly to modify and/or regulate a given sentence. These are the types of words usually representing noise in the context of IR, and normally excluded from the index via stopword lists. Open class tags, on the other hand, are, by and large, associated to the main content carriers, which are the most likely to satisfy the information need. These types of words are often morphologically productive, through inflection, conjugation, and so on, creating thus an extra hurdle to the retrieval of information. This type of problem is commonly addressed by stemming.

3 Monolingual and Bilingual Runs

The main motivation behind our participation in CLEF 2005 was to examine the performance of a set IR platform across an interesting span of lexically and morphosyntactically dissimilar languages, by revealing the extent to which retrieval models and system tuning issues are accountable for the performance of IR on a per language basis, and subsequently pay due heed to the role of language resources in the retrieval of the said languages.

We used our existing retrieval platform, which accommodates a range of matching models and a strong query expansion baseline [6]. Specifically, for the matching process, we selected the models BM25 [9], TF-IDF, as well as the

Table 1. Part-of-Speech Tagset and Class Classification

POS	Tag	Class	POS	Tag	Class
Abbreviation	ABR	Open	Ordinal Number	ORD	Closed
Adjective	JJ	Open	Possessive Ending [EN]	POS	Closed
Adverb	RB	Open	Possessive Wh-Pronoun	WP$	Closed
Auxiliary Verb	MD	Closed	Postposition [DE, GR]	POSTP	Closed
Cardinal Number	CD	Closed	Predeterminer	PDT	Closed
Conjunction	CC	Closed	Preposition	IN	Closed
Determiner	DT	Closed	Preposition with Article [FR, GR, IT]	ORD	Closed
Digit	DIG	Closed	Proclitic Noun Modifier [GR, IT]	PRN	Closed
Existentialist ”there” [EN]	EX	Closed	Pronoun	PP	Closed
Foreign Word	FW	Open	Proper Noun	NP	Open
Future Tense Particle [GR]	FUT	Closed	Quantifier [GR, IT]	QUANT	Closed
Interjection	UH	Closed	Special Preposition ”to”	TO	Closed
List Item Marker	LS	Closed	Subjunctive Particle [GR]	SUBJ	Closed
Main Verb	VV	Open	Symbol	SYM	Closed
Modal Verb	MD	Closed	Truncated Word	TR	Open
Negation Particle [FR, GR]	NEG	Closed	Wh-Adverb [EN]	WRB	Closed
Noun	NN	Open	Wh-Determiner [EN]	WDT	Closed

following Divergence from Randomness (DFR) models [1]: InexpB2, InexpC2, PL2, and DLH. For query expansion, we opted for Bo1 and Bo2 [1,8]. With the exception of the non-parametric weighting model DLH, the parameter setting of our models was realised on an empirical basis. Specifically, the matching model parameters, namely c for the DFR models, and b for BM25, were set as follows. For Bulgarian and English-Bulgarian, $c = 1.5$ and $b = 1$; for English and Greek-English, $c = 1.15$ and $b = 1$; for French and Italian-French, $c = 1$ and $b = 1$. Similarly, the query expansion $terms/documents(t/d)$ ratio was set as follows. For Bulgarian, $t/d = 25/5$; for English-Bulgarian, $t/d = 30/5$; for English and Greek-English, $t/d = 20/5$; for French, $t/d = 20/5$; for Italian-French, $t/d = 30/5$. This manifold of matching and expansion models was implemented in our Terrier retrieval platform [8].

We received our baptism of fire with Bulgarian and Greek, both of which share enough morphosyntactic and lexical complexity between them to render the need for language processing resources absolutely imperative.

Bulgarian is a Slavonic language, marked by its rich morphology and syntax, as well as by the strong lexical influence of Old Slavonic [2]. The lack of language processing resources meant that the collection was simply stemmed and indexed, without any supplementary morphosyntactic analysis. This is highly unfortunate, as even the simplest syntactic analysis could have provided the most interesting insights into the content distribution for Bulgarian. The lack of a working Bulgarian stemmer meant that stemming was realised using the Russian version of the freely available Snowball stemmer [12]. For the English - Bulgarian retrieval, the freely available Skycode machine translation system was used to translate text between the two languages [11]. The performance of the above Bulgarian and English - Bulgarian runs is summarised in Table 2. The top row relates to the topic fields used in each run, while the first column informs as

Table 2. Bulgarian and English-Bulgarian Mean Average Precision (MAP)

		Title+Description			Title+Description+Narrative		
	Model	BG	EN-BG	% mono	BG	EN-BG	% mono
Query Expansion False	BM25	0.2360	**0.1337**	56.65%	0.2174	0.1392	64.03%
	DLH	0.2211	*0.1290*	58.34%	0.2036	0.1316	64.64%
	InexpB2	0.2410	0.1266	52.53%	0.2202	0.1392	63.21%
	InexpC2	**0.2436**	0.1305	53.57%	**0.2268**	**0.1455**	64.15%
	PL2	0.2363	0.1294	54.76%	0.2203	*0.1344*	61.01%
	TF-IDF	0.2338	0.1326	56.71%	0.2173	0.1385	63.74%
Query Expansion True	BM25	**0.2662**	0.1718	64.54%	**0.2576**	**0.1864**	72.36%
	DLH	0.2409	0.1534	63.68%	0.2277	*0.1668*	73.25%
	InexpB2	0.2461	0.1538	62.49%	0.2419	0.1731	71.56%
	InexpC2	0.2618	0.1640	62.64%	0.2457	0.1846	75.13%
	PL2	*0.2514*	0.1685	67.02%	*0.2412*	*0.1799*	74.58%
	TF-IDF	0.2658	**0.1732**	65.16%	0.2574	0.1860	72.26%

to whether query expansion was used or not. BG indicates monolingual Bulgarian runs, and $EN - BG$ indicates bilingual English-Bulgarian runs. The column headed $\%mono$ relates to the difference between the monolingual and bilingual performance of corresponding runs. Submitted runs are printed in italics, and optimal runs appear in boldface.

The figures displayed in Table 2 reveal the powerful modifying influence of translation on retrieval performance, which appears to be even stronger for shorter and unexpanded topics. The overall performance of the collective matching models remains coherent throughout, as confirmed by the absence of any sharp score fluctuations. This relative stability and uniformity delineates the need for additional language processing resources for Bulgarian, the evidence of which would weigh more heavily on retrieval performance than that of simple stemming.

The second newcomer in our selection of languages was Greek, a highly inflectional Hellenic language [5]. The complexity of addressing a language as morphologically rich as Greek was accentuated by the stark lack of stemming resources. This problem received a clean treatment with the employment of a rigorous part-of-speech tagger and morphological analyser for Greek, developed by Xerox [14]. For each term in the topics, the corresponding part-of-speech and lemma was produced. When faced with two alternatives, both were selected. Closed class terms (Table 1) were rejected to reduce noise, while lemmas were automatically translated into English using Babelfish machine translation technology [3]. The performance of these runs, contrasted to their equivalent English monolingual equivalents, is presented in Table 3, in a layout similar to the one described for Table 2.

The scores presented in Table 3 are analogous to the scores relating to Bulgarian retrieval in Table 2, confirming the considerable effect of translation on the performance of the bilingual runs. Even so, the overall retrieval scores for Greek-English retrieval are significantly closer to their monolingual equivalent

Table 3. English and Greek-English Mean Average Precision (MAP)

		Title+Description			Title+Description+Narrative		
	Model	EN	GR-EN	% mono	EN	GR-EN	% mono
Query Expansion False	BM25	**0.4255**	**0.2930**	68.86%	0.4255	0.2240	52.64%
	DLH	0.4089	0.2802	68.52%	0.4089	0.2149	52.55%
	InexpB2	0.4115	0.2724	66.20%	**0.4303**	*0.2295*	53.55%
	InexpC2	0.3851	0.2758	71.62%	0.4268	**0.2386**	55.90%
	PL2	0.3634	0.2574	70.83%	0.4042	0.2126	52.60%
	TF-IDF	0.4240	0.2888	68.11%	0.4240	0.2229	52.57%
Query Expansion True	BM25	0.4556	0.3151	69.16%	0.4556	0.3151	69.16%
	DLH	0.4561	0.3128	68.58%	0.4561	0.3128	68.58%
	InexpB2	0.4307	*0.2935*	68.14%	0.4433	0.3117	70.31%
	InexpC2	0.3923	0.2678	68.49%	0.4301	0.3088	71.80%
	PL2	0.3961	0.2488	62.81%	0.4347	0.2838	65.29%
	TF-IDF	**0.4671**	**0.3168**	67.82%	**0.4671**	**0.3168**	67.82%

runs, than the overall English - Bulgarian scores are to the monolingual Bulgarian scores. This comparison underlines the auxiliary service rendered to the Greek topics by the employment of morphological analysis and lemmatisation. The performance of the bilingual Greek-English runs is in complete agreement with our primary tenet that the automatic processing of more or less recondite languages, such as Greek, cannot be entirely successful without being "aided and abetted" by some sort of morphosyntactic analysis. Stemming has been widely used in retrieval to account for this need, but it should be considered neither complete nor unique as an answer. Light syntactic analysis and lemmatisation have been shown to assist retrieval with success. Nevertheless, in order to have a measure of the relation between stemming and lemmatisation, further experimentation is needed, which would juxtapose the effect of the said methods on Greek-English retrieval.

The method used for French retrieval consisted of a variation to the monolingual French strategy tested in CLEF 2004 [6]. We opted for a less aggressive stemming approach, which targets mainly inflectional variants. Additionally, a probabilistic part-of-speech tagger [10] provided a pellucid syntactic analysis of the topics. Closed class tokens (Table 1) were removed to reduce noise. Noun phrases were extracted using the NP extractor described in the preceding section. In the case of Italian - French retrieval, Italian noun phrases were extracted and translated separately into French, using the freely available Worldlingo machine translation system [13]. The performance of the French monolingual and bilingual runs, both with the above mentioned language processing (POS - NP true) and without (POS - NP false), is presented in Table 4. Submitted runs are printed in italics, and optimal runs appear in boldface.

Table 4 reveals that the combination of part-of-speech analysis and noun phrase extraction (POS NP) is associated with better retrieval performance at all times. A point of interest is that this combination appears to benefit monolingual retrieval more than it assists bilingual retrieval. This can be deduced by the fact

Table 4. French and Italian-French Mean Average Precision (MAP)

			Title+Description			Title+Description+Narrative		
		Model	FR	IT-FR	% mono	FR	IT-FR	% mono
POS NP True	Query Expansion False	BM25	0.3199	0.2068	64.58%	0.3316	**0.2334**	70.39%
		DLH	**0.3228**	*0.2066*	64.00%	**0.3371**	0.2305	68.38%
		InexpB2	0.3171	0.2011	63.42%	0.3274	0.2245	68.57%
		InexpC2	0.3098	0.1984	64.04%	0.3198	0.2212	69.17%
		PL2	0.3092	0.2070	66.95%	*0.3206*	*0.2291*	71.46%
		TF-IDF	0.3195	**0.2073**	64.88%	0.3300	0.2328	70.54%
	Query Expansion True	BM25	0.3702	0.2763	74.63%	0.3761	0.2941	78.20%
		DLH	***0.4017***	0.2731	67.99%	**0.4198**	0.3029	72.15%
		InexpB2	0.3569	0.2444	68.48%	0.3596	0.2734	76.03%
		InexpC2	0.3480	0.2435	69.97%	0.3527	0.2676	75.87%
		PL2	*0.3765*	0.2626	69.75%	0.3809	*0.2883*	75.69%
		TF-IDF	0.3718	**0.2769**	74.47%	0.3778	**0.3045**	80.60%
POS NP False	Query Expansion False	BM25	0.3013	0.2025	67.21%	0.3083	0.2246	72.85%
		DLH	0.3007	0.1978	65.78%	0.3042	0.2184	71.79%
		InexpB2	**0.3027**	0.1976	65.28%	**0.3144**	0.2209	70.26%
		InexpC2	0.2961	0.1954	65.99%	0.3072	0.2179	70.93%
		PL2	0.2921	**0.2028**	69.43%	0.2976	0.2218	74.53%
		TF-IDF	0.3024	0.2023	66.90%	0.3087	**0.2255**	73.05%
	Query Expansion True	BM25	0.3575	0.2722	76.14%	0.3592	0.2876	80.07%
		DLH	0.3530	0.2584	73.20%	**0.3823**	**0.3015**	78.86%
		InexpB2	0.3576	0.2486	69.52%	0.3557	0.2928	82.32%
		InexpC2	0.3421	0.2425	70.88%	0.3432	0.2781	81.03%
		PL2	0.3469	0.2566	73.97%	0.3606	0.2843	78.84%
		TF-IDF	**0.3578**	**0.2748**	76.80%	0.3661	0.2989	81.64%

that the difference between the monolingual and bilingual runs is higher when POS NP is used (29.58% on average), than when it is not (26.78% on average), at all times. This observation is indicative of the fact that even though light NLP can be of significant assistance to IR, it cannot counter the shortcomings of insufficient translation resources.

The NP extractor presented above was evaluated as follows. Noun phrases were identified manually. For each noun phrase that was identified correctly by the NP extractor, a single point was added to the evaluation score. For each noun phrase that was not identified by the NP extractor, or for each non-noun phrase that was wrongly identified as a noun phrase by the NP extractor, a point was deducted. The final score of the NP extractor was compared to the manual score. Overall, the NP extractor was shown to be 88.4% accurate at identifying and extracting noun phrases for the French CLEF 2005 topic set, and 88.8% for the English. More importantly, the relation between the identification and extraction of noun phrases and the overall retrieval precision was found to be statistically significant, as per the Wilcoxon Matched-Pairs Signed-Rank Test ($p\text{-}value = 7.821e^{-10}$). Figure 1 illustrates this conclusion.

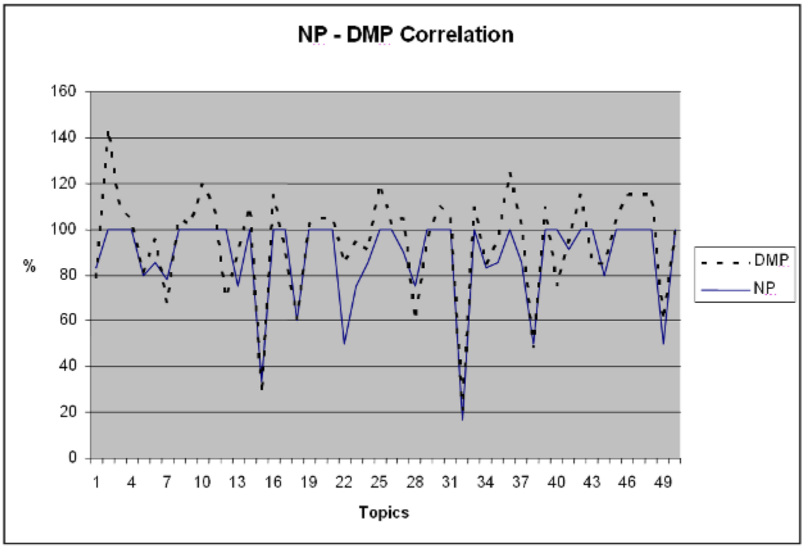

Fig. 1. Noun Phrase (NP) Extraction vs Difference from Median Precision (DMP) for French Monolingual IR

Figure 1 graphically displays the performance of the NP extractor and the difference from the median precision for our best-scoring French monolingual run as follows. The x-axis relates to the individual topics, while the y-axis relates to the percentage of the difference of firstly, the NP Extractor score from the manual score, for the NP Extractor, and secondly, the difference between the precision our best-scoring French run and the Median Precision score of all corresponding submitted runs. The said comparison throws light to the direct and strong link between the extraction of noun phrases and the overall retrieval precision, especially with respect to the median precision. Noun phrase extraction is an acknowledged procedure, and applying it to IR seems an obvious extension, without however making it the supreme arbiter. Further investigation would be required to ascertain the causal nexus between noun phrase extraction and retrieval precision.

As a conclusion, a note should be made with regards to the general performance of our retrieval platform for Bulgarian, Greek, English, French, and Italian. Figure 2 graphically plots the Mean Average Precision score (y-axis) achieved by each matching model employed for each language, or language combination, described in this paper. From the data exhibited in Figure 2, it becomes evident that runs, as clustered by language, tend to favour and disfavour specific models. Hence, DLH provides satisfying results for monolingual French retrieval. BM25, TF-IDF, and PL2 remain consistent throughout. Very frequently, the performance of all six matching models overlaps, and especially so in the case of bilingual runs. System stability aside, this trend emphasises the stultifying effect of translation on retrieval performance, as, in all cases, the overlap consists of

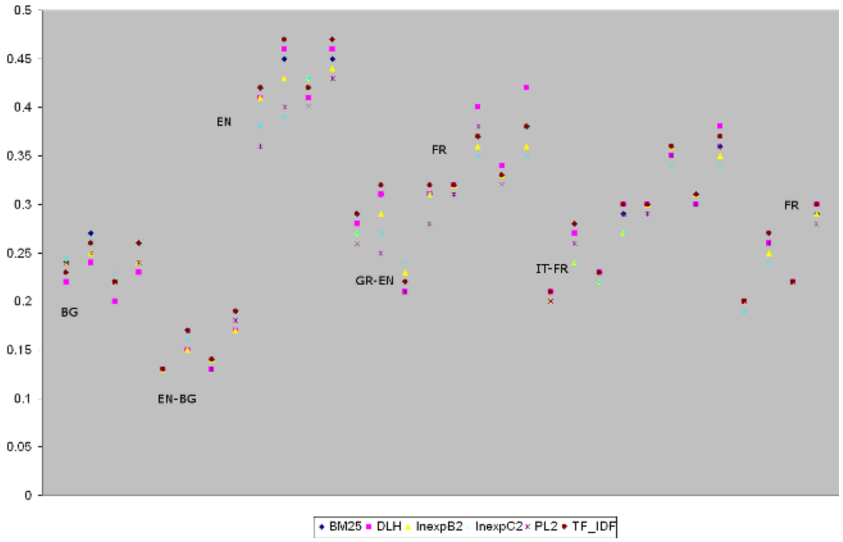

Fig. 2. Comparison of Matching Models per Language

a drop, rather than an increase of score. These results confirm the suitability of our retrieval platform for the retrieval of the aforementioned languages. In addition, they support our research scope that the poor quality and/or lack of suitable language resources for morphologically rich languages has formed an exigent set of circumstances, which cannot be addressed solely by conventional system-specific issues, such as model tuning, query expansion, and so on.

4 Conclusion

Our participation in the CLEF 2005 Ad-Hoc track for Bulgarian, English-Bulgarian, French, Italian-French, and Greek-English retrieval was shown to be successful, with a difference from the Median Precision of the collective submitted runs ranging between +1.135 (for Bulgarian) and +7.830 (for English - Greek), thus scoring second place in the English-Bulgarian and Greek-English retrieval, and third place in the monolingual French retrieval. On a collective basis, poor or no language resources were at all times associated with consistently low retrieval performance. On an individual basis, lemmatisation was shown to be a satisfactory replacement of stemming for Greek, while noun phrase extraction was shown to benefit retrieval directly and consistently for French and Italian-French. We have shown that light morphosyntactic processing can assist the retrieval of information for highly inflectional languages, and by doing so, we have carried our initial contention *a posse ad esse* successfully.

References

1. G. Amati. *Probabilistic Models for Information Retrieval based on Divergence from Randomness.* PhD thesis, Dept of Computing Science, University of Glasgow, 2003.
2. H.I. Aronson. *Bulgarian Inflectional Morphophonology.* Mouton, The Hague, 1968.
3. Babelfish Machine Translation. URL: http://babelfish.altavista.com/.
4. L. Bauer. *Introducing Linguistic Morphology.* Edinburgh University Press, 1988.
5. B. Joseph, I. Philippaki-Warburton. *Modern Greek: A Linguist's Grammar.* Croom Helm (Lingua Descriptive Series), London, 1987.
6. C. Lioma, B. He, V. Plachouras and I. Ounis. The University of Glasgow at CLEF 2004: French Monolingual Information Retrieval with Terrier. In Peters, C., Clough, P. D., Jones, G. F. J., Gonzalo, J., Kluck, M., Magnini, B. (eds.): *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign.* Lecture Notes in Computer Science, Springer-Verlag, 2005.
7. M.P. Marcus, B. Santorini and M.A. Marcinkiewicz. Building a Large Annotated Corpus for English: The Penn Treebank. In *Computational Linguistics*, Volume 19, Number 2, pp. 313–330, 1993.
8. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier Information Retrieval Platform. In *Proceedings of ECIR 2005*, LNCS vol. 3408, pp. 517–519, 2005. URL: http://ir.dcs.gla.ac.uk/terrier/.
9. S.E. Robertson. Okapi at TREC-3. In Harman, D. K. (eds.): *Overview of the Third Text Retrieval Conference (TREC-3)*, NIST, 2005.
10. H. Schmidt. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Jones, D., Somers, H. (eds.): *New Methods in Language Processing Studies.* Computational Linguistics, UCL Press, 1997.
11. Skycode Machine Translation. URL: http://webtrance.skycode.com/online.asp/
12. Snowball stemmers. URL: http://snowball.tartarus.org/.
13. Worldlingo Machine Translation. URL: http://www.worldlingo.com/.
14. Xerox Greek Language Analysis. URL: http://www/xrce.xerox.com/competencies/content-analysis/demos/greek/