Overview of the CLEF 2005 Multilingual Question Answering Track

Alessandro Vallin [*], Bernardo Magnini [†], Danilo Giampiccolo [‡], Lili Aunimo [§], Christelle Ayache [¶], Petya Osenova [‖], Anselmo Peas [**], Maarten de Rijke [††], Bogdan Sacaleanu [‡‡], Diana Santos [#], Richard Sutcliffe [&]

September 9, 2005

Abstract

The general aim of the third CLEF Multilingual Question Answering Track was to set up a common and replicable evaluation framework to test both monolingual and cross-language Question Answering (QA) systems that process queries and documents in several European languages. Nine target languages and ten source languages were exploited to enact 8 monolingual and 73 cross-language tasks. Twenty-four groups participated in the exercise.Overall results showed a general increase in performance in comparison to last year. The best performing monolingual system irrespective of target language answered 64.5% of the questions correctly (in the monolingual Portuguese task), while the average of the best performances for each target language was 42.6%. The cross-language step instead entailed a considerable drop in performance. In addition to accuracy, the organisers also measured the relation between the correctness of an answer and a system's stated confidence in it, showing that the best systems did not always provide the most reliable confidence score.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering

[*] CELCT, Italy,{vallin}@celct.it.

[†] ITC-Irst, Italy, magnini@itc.it.

[‡] CELCT, Italy,{giampiccolo}@celct.it.

[§] University of Helsinki, Finnland, aunimo@cs.helsinki.fi.

[¶] ELDA/ELRA, France, ayache@elda.org.

[‖] BTB, Bulgaria, petya@bultreebank.org.

[**] UNED, Spain, anselmo@lsi.uned.es.

[††] University of Amsterdam, The Netherlands, mdr@science.uva.nl.

[‡‡] DFKI, Germany, Bogdan.Sacaleanu@dfki.de

[#] *Sintef, Norway,* Diana.Santos@sintef.no.

[&] *UniversityofLimerick, Ireland,* Richard.Sutcliffe@ul.ie.

1 Introduction

The CLEF QA evaluation campaign conducted in 2005 [1] was the result of the experience acquired during the two previous campaigns and of the proposals suggested in last year's workshop in order to make the track more challenging and realistic.

At a first look one realizes that over the years the series of QA competitions at CLEF has registered a steady increment in the number of participants and languages involved, which is particularly encouraging as multilinguality is one of the main characteristic of these exercises. In fact, in the first campaign, which took place in 2003, eight groups from Europe and North America participated in nine tasks, three monolingual -Dutch, Italian and Spanish- and five bilingual, where questions were formulated in five source languages -Dutch, French, German, Italian- and answer were searched in an English corpus collection. In 2004 eighteen groups took part to the competition, submitting 48 runs. Nine source languages -Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese and Spanish- and 7 target languages -all the source languages but Bulgarian and Finnish, which had no corpus available- were considered in the task. In 2005 the number of participants rose to twenty-four, 67 runs were submitted, and 10 source languages -the same as those used in the previous year plus Indonesian- and 9 source languages -the same used as sources, except Indonesian which had no corpus available- were exploit in 8 monolingual and seventy-three cross-language tasks. Moreover, some innovation was introduced concerning the type of questions proposed in the exercise and the metrics used in the evaluation. Despite the expectation of some the organisers for more radical innovations were partially disappointed, this edition of QA@CLEF was altogether successful and can be considered a good starting point for next campaigns.

A preliminary overview of the 2005 QA track is presented in this paper, explaining more in details the procedure followed to build the test sets and providing a preliminary analysis of the results.

2 Tasks

The tasks proposed in 2005 QA campaign were characterised by a basic continuity with what had been done in 2004. In fact, to the demand for more radical innovation a more conservative approach was preferred, as most organizers opted to further investigate the procedures consolidated in the last two campaigns before moving to the next stage. As a matter of fact, the task remained basically the same as that proposed in 2005, although some minor changes were actually introduced, i.e. a new type of questions, and two new measures, namely K1 measure and r value. Both question type and measure were borrowed from the Spanish pilot task proposed at CLEF 2004 [2].

2.1 Languages

Ten source languages -Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese, Spanish and, as an experiment, Indonesian- and 9 target languages -all the source languages except Indonesian- were considered at the 2005 CLEF QA track. Eighty-one tasks were setup, 8 monolingual -Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese, Spanish- and 73 bilingual. In this way, all the possible combinations between source and target languages were exploited, but for two exceptions: Indonesian, being included in a cross-language QA competition, was used only as a source in the Indonesian-English task, meanwhile the monolingual English task was discarded as it has been abundantly tested in TREC campaigns, according to the decision taken in the previous competition.

[1] For more information about QA@CLEF campaigns visit *http://clef-qa.itc.it.*

2.2 The Evaluation Exercise

As in the previous campaign, 200 questions were provided as an input in all tasks, and exact answer-strings were required as an output. The target corpora in all the languages were collections of newspapers and news agencies' articles, whose texts had been SGML tagged. Each document had a unique identifier (docid) that systems had to return together with the answer, in order to support it. The corpora, released by ELRA/ELDA, were large, unstructured, open-domain text collections.
Although the numbers of questions was the same as last year, there were changes regarding the type of questions and their distribution. Meanwhile How- and Object questions were not included in 2005 task, since they were considered particularly problematic in the evaluation phase, a new subtype of factoid questions was introduced, called *temporally restricted* questions, which constrained by either an event -e.g. *Who was Uganda's President during Rwanda's war?*-, a date -e.g. *Which Formula 1 team won the Hungarian Grand Prix in 2004?*- or a period of time-e.g. *Who was the President of the European Commission from 1985 to 1995?*. Up to 30 temporally restricted questions could be included in each task.
As far as the three major type of questions -Factoids (F), Definition (D) and NIL (N)- are concerned, the breakdown, both suggested and real, is shown in Table 1.

Table 1: **Test set breakdown according to question type**

	F	**D**	**T**	**NIL**
suggested	[120]	[50]	[30]	[20]
BG	116	50	34	22
DE	135	42	23	20
EN	121	50	29	20
ES	118	50	32	20
FI	111	60	29	20
FR	120	50	30	20
IT	120	50	30	20
NL	114	60	26	20
PT	135	42	23	18

In order to increase the overlap between the test sets of different target languages, this year a certain number of topics were assigned to each language and a particular effort was made in order to get general questions, which could easily find an answer also in the other corpora. As a result, no question was actually answered in all 9 languages, but the inter-language partial overlap was increased anyway, as 21 questions appeared in 5 target languages, 66 questions in 4 target languages, 159 questions in 3 target languages and 265 in 2 target languages.

3 Test Set Preparation

The procedure for question generation was the same as that adopted in the previous campaigns. Nine groups were involved in the generation, translation and manual verification of the questions: the Bulgarian Academy of Science, Sofia, Bulgaria (CLPP) was in charge for Bulgarian; the Deutsches Forschungszentrum fr Knstliche Intelligenz Saarbrcken, Germany, (DFKI) for German; the Evaluations and Language Resources Distribution Agency Paris, France(ELRA/ELDA) for French; the Center for the Evaluation of Language and Communication Technologies Trento, Italy (CELCT) for Italian; Linguateca, Oslo (Norway), Braga, Lisbon& Porto for Portuguese; the Universidad Nacional de Educacin a Distancia Madrid, Spain (UNED) for Spanish, the University of Amsterdam, The Netherland for Dutch; the University of Helsinki, Finland for Finnish; the University of Limerick, Ireland for English; and the Department of Computer Science of University of Indonesia joined the activity translating 200 English questions into Indonesian, in order to set up the cross-language Indonesian- English task.

3.1 Question Generation

The questions in the test sets addressed large open domain corpora, mostly represented by the same comparable document collections used last year: *NRC Handelsblad* (years 1994 and 1995) and *Algemeen Dagblad* (1994 and 1995) for Dutch; *Los Angeles Times* (1994) and *Glasgow Herald* (1995) for English; *Le Monde* (1994) and *SDA French* (1994 and 1995) for French; *Frankfurter Rundschau* (1994), *Der Spiegel*(1994 and 1995) and *SDA German* (1994 and 1995) for German; *La Stampa* (1994) and *SDA Italian* (1994 and 1995) for Italian; *PUBLICO*(1994 and 1995) for Portuguese and *EFE* (1994 and 1995) for Spanish. This year two new corpora were added, *Aamulehti* (1994-1995) for Finnish, and *Sega* and *Standard* for Bulgarian. Unfortunately the Bulgarian corpora dated back to 2002, so that the information contained in it was difficulty comparable with the that of the other corpora.

According to the consolidate procedure, 100 questions were produced in each target language (except Indonesian), manually searching relevant documents for at least one answer. The questions were then translated into English, so that could be understood and reused by all the other groups. Answers were not translated this year, as it was a time-consuming and basically useless activity.

The co-ordinators attempted to balance the difficulty the test sets according to the different answer types of the questions already used in the previous campaigns, i.e. TIME , MEASURE, PERSON, ORGANISATION, LOCATION, and OTHER. HOW and OBJECT questions were however inserted in this exercise because generate ambiguous responses, which are quite difficult to be assessed.

As said, up to thirty *temporally restricted* questions were allowed, and were themselves classified according to the above mentioned types, ie, time, measure, etc Particular care was taken this year in choosing 10% of NIL questions. In fact, some organizers realised that in the previous campaigns NIL questions were quite easily identified by systems, as they were manually generated searching for named entities which were not in the corpora. On the contrary, this time NIL questions were selected randomly from those that seemed to have no answer in the document collections, and were double-checked Once the 900 questions were formulated in the original source languages, translated into English and collected in a common XML format, native speakers of each source language, with a good command of English were recruited to translate the English version of all the other questions trying to adhere as much as possible to the original. This process was as challenging as any translation job can be, since many cultural discrepancies and misunderstanding easily creep in. Anyway, as was already pointed out in 2004 "he fact that manual translation captured some of the cross-cultural as well as cross-language problems is good since QA systems are designed to work in the real world" [3].

3.2 Gold Standard

Once all the 900 questions were translated into ten source languages -the Indonesian group translated only the final 200 English question-, 100 additional questions for each target language were selected from the other source languages, so that at the end each language had 200 questions. The added questions were manually verified and searched for answers in the corpus of the respective language. The collection was called *Multi9-05*, and was presented in the same XML format adopted in 2004. The entire collection is made up of 205 definition questions and 695 factoid , which are quite well balanced according to their types, being divided as follows: 110 MEASURE; 154 PERSON; 136 LOCATION; 103 ORGANISATION, 107 OTHER, 85 TIME. The total number of temporally restricted questions is 149.

Although this new kind of questions appeared to be quite interesting, no comprehensive analysis of the results in this group of questions have been made so far, and the experiment requires to be furtherly investigated. The *Multi9-05* can now be added to the previous campaigns' collections, which already represent a useful reusable benchmark resource. The proposal to integrate the missing answers with the correct results provided by the systems during the exercise has remained unanswered.

3.3 Participants

The positive trend in terms of participation registered in 2004 was confirmed in the last campaign. From the original 8 groups who participated in 2003 QA task, submitting a total of 19 runs in 9 tasks, the number of competitors has raised to twenty-four, which represent an increase of 33% respect to last year, when 18 groups took part in the exercise. The total of submitted runs was sixty-seven. All the participants in 2005 competition were from Europe, with the exception of group from University of Indonesia which tried the experimental cross-language task Indonesian-English.

Table 2: Runs and Participants

	BG$_t$		**DE**$_t$		**EN**$_t$		**ES**$_t$		**FI**$_t$		**FR**$_t$		**IT**$_t$		**NL**$_t$		**PT**$_t$	
	R	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**	**P**
BG$_s$	**2**	**2**	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-	-
DE$_s$	-	-	**3**	**2**	1	1	-	-	-	-	-	-	-	-	-	-	-	-
EN$_s$	-	-	3	2			3	2	-	-	1	1	-	-	-	-	1	1
ES$_s$	-	-	-	-	1	1	**13**	**7**	-	-	-	-	-	-	-	-	-	-
FI$_s$	-	-	-	-	2	1	-	-	**2**	**1**	-	-	-	-	-	-	-	-
FR$_s$	-	-	-	-	4	2	-	-	-	-	**10**	**7**	-	-	-	-	-	-
IN$_s$					1	1												
IT$_s$	-	-	-	-	2	1	2	1	-	-	1	1	**6**	**3**	-	-	-	-
NL$_s$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**3**	**2**	-	-
PT$_s$	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-	**4**	**3**

As shown in table 2, the systems were tested only against 22 of the 81 activated tasks. Monolingual English was discarded this year, as it was in last competition, because the task has been sufficiently investigated in TREC campaigns, and as far as Indonesian is concerned, only the task with English as a target was set up. The non-activated tasks are represented by a blank cell in table 2.

All nine monolingual tasks (in bold in the table) were tested at least by 1 systems, being French (FR) and Spanish (ES) the most chosen languages. As far as bilingual tasks are concerned, 15 participants altogether chose to test their systems in a cross-language task. English was as usual the most frequent target language, being involved in 8 cross-lingual tasks completed by 9 participants; Spanish was chosen as a target in a cross-language task by three groups, and so was French, meanwhile only one system tried a cross-language task with Portuguese (PT) as a target, i.e. EN-PT. All the other languages was not considered as a target in bilingual tasks.

4 Results

The procedure adopted to assess the systems's outputs was practically the same as the last year. Participants were allowed to submit just one response per question and up to two runs per task, which were judged by human assessors according to correctness and exactness -where correctness expresses whether the answer is clear and pertinent, while exactness evaluates whether the information is either too much or too less. Like in 2004 only exact answers were allowed, and the responses were judged as Right, Wrong, ineXact or Unsupported (when the answer-string contained a correct answer but the returned docid did not support it). As a partial analysis of the inter-tagger agreement has shown, the exactness is still a major problem in evaluation, as most disagreement between judges concerns this parameter.

Definition questions, which were introduced last years, and were considered particularly difficult also because they could raise problems in assessing their exactness, generally scored quite well, proving that as they are now they are less challenging than one thought. In fact, the answer often consists in the solution of an acronym, when they concern organisation, or is expressed as an apposition of the proper name, when persons are concerned. As said, the introduction of Temporal Restricted Questions hasn't been properly analysed yet. As a general remark, it must be said that their number in the test sets was probably too small to provide significative data on their impact on systems' results. Furthermore, some of them were "false temporally restricted", as said above, and a system could retrieve an answer without even considering the temporal restriction.

The main measure used for the evaluation was the accuracy, i.e the fraction of right answers. The answers were returned unranked (i.e. in the same order as in the test set), but a confidence value, that could range between 0 and 1, could be added to each string and be considered to calculate an additional Confidence-weighted Score (CWS)[NOTA OVERVIEW 2004]. This year two additional evaluation measures, i.e. the K1 value and r coefficient, borrowed by [VEDERE NOTA], were experimentally introduced, in order to find a comprehensive measure which takes into account both accuracy and confidence. Anyway, being confidence an additional and optional value, only some systems could be assigned the CWS, and consequently the K1 and r coefficient; therefore an analysis based on this measures is not very significant at the moment.

In comparison to last year, the performances of the systems in this campaign show a general improvement, although a significant variation remains among target languages. In fact, in 2004 the best performing monolingual system irrespective of target language (henceforth 'best overall') answered 45.5% of the questions correctly, while the average of the best performances for each target language (henceforth 'average of best') was 32.1%. In 2005 the best overall and average of best figures were 64.5% (in the monolingual Portuguese task)-representing an increase of 19 point- and 42.6% respectively. As far as bilingual tasks are concerned, as usual the cross-lingual step generically entailed a considerable drop in performance.

In the following nine sections the results of the runs for each target language are thoroughly discussed. For each target language two kinds of results are given, summarized in two tables. One presents the overall performance, giving the number of right (R), wrong (W), inexact (X), and unsupported (U) answers; the accuracy, in general and on Factoids (F), Definitions (D) and Temporal (T); Precision (P), Recall (R) and F measure for NIL questions; and finally CWS, K1 and r of each run. The second table shows the accuracy of the systems with respect to the answer types, i.e Definition, sub-classified as Organisation (Or) and Person (Pe), and Factoid and Temporally Restricted, sub-classified as location (Lo), measure (Me), organisation (Or), other (Ot), person (Pe) and time (Ti). Below each answer type, the number of posed questions of that type is shown in square brackets.

The last row of the second table shows a virtual run, called Combination, in which the classification "right answer" is assigned to a question if any of the participating systems found it. The objective of this combination run is to show the potential achievement if one merged all answers and considered the set of answers right, provided one answer were right.

4.1 Bulgarian as Target

For the first time Bulgarian was addressed as a target language at CLEF 2005. Thus, no comparison can be made with previous results from the same task, but some comments on the present ones are in order.

This year two groups participated in monolingual evaluation tasks with Bulgarian as a target language: IRST, Trento and BTB, LML, IPP, Sofia. Two runs were submitted for Bulgarian-Bulgarian. Both results are below the desired figures (27.50 % and 18.50 % correct answers), but they outperform their own results from the last year where Bulgarian was used as a source language and English - as a target. Obviously, the Inexact and Unsupported value metrics do not have substantial impact over the final estimations. It seems that as a group the definition questions are the best assessed type (40 % and 42 %). Then come the factoid ones. The worst performance goes to the temporally restricted questions. Then, NIL questions exhibit better recall than precision. It might be explained by the fact that the systems return NIL when they are not sure in the answer. Only IRST group results provide a confidence weighted score.

It is interesting to discuss the results according to the answer types. Recall that definitions did well as a group. However, when divided further into Organization and Person types, it turns out that the Organization type was better handled by one of the participants, while the Person type was better handled by the other. From non-temporally restricted factoids Organizations and

Table 3: **Results in the tasks with Bulgarian as target**

run	**Right**		**W**	**X**	**U**	**Right**			**NIL** [22]			**CWS**	**K1**	**r**
	#	%	#	#	#	**% F** [116]	**% D** [50]	**% T** [34]	**P**	**R**	**F**			
irst051bgbg$_{\mathrm{M}}$	55	27.50	130	13	2	25.00	40.00	17.65	0.15	0.41	0.22	0.144	-0.035	0.160
btb051bgbg$_{\mathrm{M}}$	37	18.50	160	3	-	10.34	42.00	11.76	0.05	0.41	0.10	-	-	-

Other have been the most problematic types. From temporally restricted factoids Measure was unrecognized, but the number of these questions was not so high anyway. Person subtype was not detected as well, which is a bit surprising fact.

Table 4: **Results in the tasks with Bulgarian as target (breakdown according to answer type**

run	*Correct Answers*														
	Definition		*Factoid*						*Temporally restricted factoid*					*Total*	
	Or [25]	**Pe** [25]	**Lo** [19]	**Me** [20]	**Or** [18]	**Ot** [19]	**Pe** [20]	**Ti** [20]	**Lo** [7]	**Me** [7]	**Or** [4]	**Ot** [4]	**Pe** [12]	#	%
irst051bgbg$_{\mathrm{M}}$	6	14	6	4	2	3	7	7	-	3	2	-	-	55	27.50
btb051bgbg$_{\mathrm{M}}$	13	8	2	2	1	2	2	3	2	-	2	-	-	37	18.50
combination	**16**	**17**	**6**	**4**	**2**	**3**	**7**	**9**	**2**	**3**	**2**	**-**	**1**	**72**	**36.00**

Most of the problems concerning assessors' agreement were in one 'green area': between Wrong and Inexact. Recall that it was also a problem at CLEF 2004. Here we do not have in mind easy cases, such as: What is FARC? The system answered 'Columbia' instead of answering 'Revolutionary Armed Forces of Colombia' or at least 'Revolutionary Armed Forces'. We have in mind subtle cases as follows: (1) too general answers, but still correct (Q: What is ESA? A: 'agency' instead of '(European) space agency'), and (2) partial answers, but still correct (Q: Who was proclaimed patron of Europe by the Pope on 31 December 1980? A: 'St. Cyril' instead of 'St. Cyril and Methodius'). Under the former type we consider answers that are given only some 'top ontological' categorization. Under the latter we consider cases, in which part of the answer is presented, but the other part is missing. Very often it concerns questions of measure (Q: How much did Greenpeace earn in 1999? A: '134' instead of '$ 134 mln.').
This year for the first time Bulgarian was tested as a target language at the CLEF track. Two groups made runs on Bulgarian-Bulgarian task. The results are promising in spite of being lower than the half of the correctly recognized answers. So, we consider this a good start. The two extraction systems will be improved on the evaluation feedback. They need to handle better local contexts as well as to try to handle non-local support information.
In the evaluation phase the most problematic still seems to be the definition of the Inexact answer. Inexactness exhibit gradability. In this respect it either should be defined in a more elaborate way (concerning generality and partiality, and per answer type), or there should be introduced a more objective system of final evaluation. Our suggestion is that inexact answers have to contain the head noun of the correct answer. The degree of inexactness depends on the recognized modifiers of the head. If the correct answer is a coordination, then the inexactness is determined also by presence of each coordinates.

4.2 German as Target

There were three research groups that took part in this year's evaluation for the QA-track having German as target language. The number of total system runs submitted by the participants was six, with three runs for every of the two source languages: German and English. The results of evaluation for every participant group are shown in the tables below.

For the monolingual German runs the results for definition and temporal questions are better then those for factoid questions. As table 6 shows, within the definition questions, results are

Table 5: **Results in the tasks with German as target**

run	**Right**		**W**	**X**	**U**	**Right**			**NIL** [20]			**CWS**	**K1**	**r**
	#	%	#	#	#	**% F** [135]	**% D** [42]	**% T** [23]	**P**	**R**	**F**			
dfki051dede$_{\mathbf{M}}$	87	43.50	100	13	-	35.83	66.00	36.67	0.29	0.65	0.40	0.385	0.095	0.300
fuha051dede$_{\mathbf{M}}$	72	36.00	119	9	-	25.00	70.00	23.33	0.14	1.00	0.25	0.346	0.221	0.665
dfki052dede$_{\mathbf{M}}$	54	27.00	127	19	-	15.00	52.00	33.33	0.28	0.65	0.39	0.227	0.045	0.386
dfki051ende$_{\mathbf{C}}$	46	23.00	141	12	1	16.67	50.00	3.33	0.09	0.10	0.09	0.201	0.060	0.483
dfki052ende$_{\mathbf{C}}$	31	15.50	159	8	2	8.33	42.00	0.00	0.08	0.10	0.09	0.137	0.040	0.564
uhiq051ende$_{\mathbf{C}}$	10	5.00	161	29	-	0.83	18.00	0.00	0.00	0.00	0.00	0.006	-0.310	0.080

better for ORGANIZATION as for PERSON answer types. For factoid questions, best results were attained for TIME, PERSON, LOCATION and ORGANIZATION answer types, in order of their mention, while for temporal questions, results were equally good for PERSON, MEASURE and ORGANIZATION answer types.

For the cross-lingual English-German runs, best results were registered for definition questions, followed by factoid questions, and with poor results by temporal questions. Again, best results for definition questions were for ORGANIZATION answer types and for factoid questions the order of accuracy remains unchanged with respect to the monolingual runs.

Results computed for a "virtual" system, through aggregation of all existing results, show an increase of almost 35% for the monolingual task, and 20% for the cross-lingual task, in accuracy over the best results achieved by participating systems.

Table 6: **Results in the tasks with German as target (breakdown according to answer type)**

run	***Correct Answers***														
	Definition		***Factoid***						***Temporally restricted factoid***					***Total***	
	Or [29]	**Pe** [21]	**Lo** [21]	**Me** [20]	**Or** [20]	**Ot** [19]	**Pe** [20]	**Ti** [20]	**Lo** [4]	**Me** [13]	**Or** [3]	**Ot** [3]	**Pe** [7]	#	%
dfki051dede$_{\mathbf{M}}$	22	11	7	6	5	3	8	14	2	5	1	-	3	87	43.50
fuha051dede$_{\mathbf{M}}$	20	15	5	2	3	8	5	7	-	4	1	-	2	72	36.00
dfki052dede$_{\mathbf{M}}$	20	6	2	-	3	4	3	6	2	4	1	-	3	54	27.00
combination	**28**	**17**	**10**	**8**	**6**	**10**	**9**	**16**	**2**	**7**	**1**	**-**	**3**	**117**	**58.50**
dfki051ende$_{\mathbf{C}}$	21	4	6	1	-	1	4	8	-	1	-	-	-	46	23.00
dfki052ende$_{\mathbf{C}}$	20	1	2	-	1	1	4	2	-	-	-	-	-	31	15.50
uhiq051ende$_{\mathbf{C}}$	3	6	-	-	-	-	1	-	-	-	-	-	-	10	5.00
combination	**22**	**8**	**7**	**1**	**1**	**2**	**6**	**8**	**-**	**1**	**-**	**-**	**-**	**56**	**28**

4.3 English as Target

Overall, twelve cross-lingual runs with English as a target were submitted. The results are shown in Tables 7 and 8.

The best scoring system overall was DFKI DEEN Run 1 with 25.5%. This score includes all three types of question, i.e. Factoid, Definition and Temporal. For Factoid questions alone, the highest scoring was DLTG FREN Run 1 (20.66%). For Definition questions alone, the highest scoring was DFKI DEEN Run 1 (50%). For Temporal question alone, three systems had an equal top score, DLTG FREN Run 2, IRST BGEN Run 1 and LIRE FREN Run 2 (all 20.69%). DFKI's main advantage over other systems was their ability to answer definition questions - their score of 50% was well ahead of the next best score of 38% achieved by IRST ITEN Run 1 and IRST ITEN Run 2.

Last year, results were only single-judged with all answers to a given question being judged by one assessor using an adapted version of the NIST software. Four assessors each did 50 questions, there being 200 in all. Any issues found by assessors were then discussed and resolved at a series of plenary sessions. This year, all results were double-judged using the same software and with six

Table 7: **Results in the tasks with English as target**

run	**Right**		**W**	**X**	**U**	**Right**			**NIL** [20]			**CWS**	**K1**	**r**
	#	%	#	#	#	**% F** [121]	**% D** [50]	**% T** [29]	**P**	**R**	**F**			
dfki051deen$_C$	51	25.50	141	8	-	18.18	50.00	13.79	0.22	0.50	0.31	0.203	0.000	0.322
irst051iten$_C$	47	23.50	145	6	2	19.83	38.00	13.79	0.18	0.35	0.24	0.118	-0.141	0.240
lire052fren$_C$	38	19.00	152	9	1	16.53	24.00	20.69	0.24	0.45	0.31	0.048	-0.201	0.088
irst051bgen$_C$	37	18.50	145	17	1	17.36	20.00	20.69	0.17	0.35	0.23	0.079	-0.270	0.055
dltg051fren$_C$	36	18.00	149	15	-	20.66	12.00	17.24	0.11	0.30	0.16	-	-	-
dltg052fren$_C$	36	18.00	151	13	-	19.83	12.00	20.69	0.10	0.30	0.14	-	-	-
upv051esen$_C$	34	17.00	156	9	1	12.40	28.00	17.24	0.15	0.50	0.23	0.072	-0.105	0.152
lire051fren$_C$	28	14.00	156	14	2	13.22	18.00	10.34	0.21	0.15	0.18	0.043	-0.225	0.237
irst052iten$_C$	26	13.00	168	6	-	5.79	38.00	-	0.22	0.50	0.31	0.114	-0.328	0.414
hels051fien$_C$	25	12.50	164	10	1	12.40	12.00	13.79	0.17	0.55	0.27	0.050	-0.338	0.022
hels052fien$_C$	20	10.00	167	11	2	10.74	8.00	10.34	0.21	0.40	0.27	0.041	-0.332	0.058
uixx051inen$_C$	2	1.00	162	36	-	-	4.00	-	0.40	0.10	0.16	8e-05	-0.770	0.253

assessors: Two independently judged questions 1-66, two judged 67-133 and two judged 134-200, there being 200 questions in total once again. The judgements were then automatically compared using the diff utility. A list of variant judgements was then prepared and presented to each pair of assessors for resolution.

Table 8: **Results in the tasks with English as target (breakdown according to answer type)**

run	***Correct Answers***														
	Definition		*Factoid*						*Temporally restricted factoid*					*Total*	
	Or [25]	**Pe** [25]	**Lo** [20]	**Me** [20]	**Or** [20]	**Ot** [21]	**Pe** [20]	**Ti** [20]	**Lo** [2]	**Me** [9]	**Or** [3]	**Ot** [5]	**Pe** [10]	#	%
dfki051deen$_C$	12	13	4	4	5	1	2	6	-	2	-	-	2	51	25.50
irst051iten$_C$	4	15	7	1	3	3	2	8	-	1	-	1	2	47	23.50
lire052fren$_C$	5	7	6	4	4	1	-	5	-	2	1	2	1	38	19.00
irst051bgen$_C$	6	4	7	4	4	1	3	2	-	2	1	-	3	37	18.50
dltg051fren$_C$	2	4	8	4	1	2	2	8	1	2	-	1	1	36	18.00
dltg052fren$_C$	3	3	8	4	1	2	1	8	1	3	-	1	1	36	18.00
upv051esen$_C$	7	7	2	3	3	1	-	6	-	4	-	1	-	34	17.00
lire051fren$_C$	5	4	7	2	1	1	-	5	-	-	2	-	1	28	14.00
irst052iten$_C$	4	15	-	2	2	-	-	3	-	-	-	-	-	26	13.00
hels051fien$_C$	6	-	3	1	2	2	2	5	-	2	-	2	-	25	12.50
hels052fien$_C$	4	-	1	1	2	1	2	6	-	1	-	2	-	20	10.00
uixx051inen$_C$	2	-	-	-	-	-	-	-	-	-	-	-	-	2	1.00
combination	**19**	**23**	**17**	**11**	**10**	**6**	**6**	**13**	**1**	**6**	**2**	**4**	**5**	**123**	**61.50**

The degree of agreement between assessors was found to range between 91.41% and 94.90%, computed as follows: For questions 1-66 there were 66 questions and 12 runs, 792 judgements in all. 68 differences were recorded, so the level of agreement is (792-68)/792, i.e. 91.41%. For questions 67-133, there were 804 jusgements with 69 differences recorded, i.e. 91.42% agreement. Finally, for questions 134-200 there were again 804 judgements with 41 differences recorded, i.e. 94.90% agreement.

In almost all cases, points of disagreement could be tracked down to problematic questions which either had no clear answer (but several vague ones) or which had several possible answers depending on the interpretation of the question.

Definition questions were once again included this year but a method of assessing them was not decided upon prior to the competition. In other words, participants did not really know what sort of system to build for definitions and we as assessors were unsure how to go about judging the answers. In consequence we used the same approach as last year: If an answer contained information relevant to the question and also contained no irrelevant information, it was judged R if supported, and U otherwise. If both relevant and irrelevant information was present it was judged X. Finally, if no relevant information was present, the answer was judged W. Two main types of system were used by participants, those which attempted to return an exact factoid-style

answer to a question, and those which returned one or more text passages from documents in the collection. Generally, the former type of system is attempting a harder task because it is returning more concise information than is the latter type of system. For this reason, our evaluation method is designed to favour the former type. This was an arbitrary decision, taken in the absence of further guidelines. Our judgements are as accurate as we can make them within our own criteria but we should point out that different criteria could produce different results.
Concerning the overall assessment process, we had no procedural difficulties as the format of the data was the same as last year and Michael Mulcahy in particular had already devoted a great deal of time to the adaptation of the software and the development of additional utilities in 2004. Also, most of the assessors were familiar both with the software and with the judgement criteria. We arrived at two conclusions during the assessment process. Firstly, the main points of difference between assessors in judging answers can be traced back to intrinsic problems associated with certain questions. In other words we need to devote more time to the problem of generating good questions which on the one hand are of the kind which potential users of our systems might pose, and on the other hand have clear answers. We should arrive at objective tests which can be applied to a candidate question and its answers to enable its suitability for use in CLEF to be assessed. Secondly, the situation in respect of definition questions was not ideal for either participants or assessors. This could affect our results for the EN target language as well as their relationship to the results for other target languages.

4.4 Spanish as Target

Seven groups submitted 18 runs having Spanish as target language: 13 of them had also Spanish as source language, 2 had Italian and 3 had English. Notice that is the first time that bilingual runs were submitted.

Table 9 shows the number of correct answers, CWS, *K1* and correlation coefficient for all systems. Table 10 shows the number of correct answers for each type of question. Table 11 shows the number of correct answers for each type of temporal restriction. Table 12 shows the evolution of the most important criteria in the systems performance for the last three years.

Table 9: **Results in the tasks with Spanish as target**

run	**Right**		**W**	**X**	**U**	**Right**			**NIL** [20]			**CWS**	**K1**	**r**
	#	**%**	**#**	**#**	**#**	**% F** [118]	**% D** [50]	**% T** [32]	**P**	**R**	**F**			
inao051eses$_\mathbf{M}$	84	42.00	110	5	1	28.81	80.00	31.25	0.23	0.80	0.36	-	-	-
tova051eses$_\mathbf{M}$	82	41.00	109	7	2	28.81	80.00	25.00	0.24	0.55	0.33	-	-	-
inao052eses$_\mathbf{M}$	79	39.50	116	4	1	27.12	80.00	21.88	0.19	0.80	0.31	-	-	-
tova052eses$_\mathbf{M}$	77	38.50	113	8	2	23.73	80.00	28.12	0.22	0.55	0.32	-	-	-
upv051eses$_\mathbf{M}$	67	33.50	119	13	1	26.27	52.00	31.25	0.19	0.30	0.23	0.218	0.043	0.338
alia051eses$_\mathbf{M}$	66	33.00	110	24	-	29.66	40.00	34.38	0.25	0.45	0.32	0.170	-0.273	0.038
aliv051eses$_\mathbf{M}$	65	32.50	116	18	1	28.81	46.00	25.00	0.26	0.25	0.26	0.15	-0.224	0.223
alia052eses$_\mathbf{M}$	60	30.00	114	26	-	26.27	36.00	34.38	0.24	0.45	0.32	0.153	-0.323	0.038
talp051eses$_\mathbf{M}$	58	29.00	122	20	-	27.97	36.00	21.88	0.26	0.70	0.38	0.089	-0.185	-0.011
talp052eses$_\mathbf{M}$	54	27.00	133	13	-	25.42	32.00	25.00	0.22	0.65	0.33	0.078	-0.210	-0.043
mira051eses$_\mathbf{M}$	51	25.50	138	11	-	26.27	34.00	9.38	0.08	0.10	0.09	0.123	-0.302	0.315
mira052eses$_\mathbf{M}$	46	23.00	140	14	-	22.03	34.00	9.38	0.08	0.10	0.09	0.103	-0.343	0.316
upv052eses$_\mathbf{M}$	36	18.00	155	9	-	22.88	0.00	28.12	0.10	0.40	0.16	0.128	0.041	0.563
upv051enes$_\mathbf{C}$	45	22.50	139	14	2	19.49	34.00	15.62	0.15	0.20	0.17	0.103	-0.033	0.197
mira052enes$_\mathbf{C}$	39	19.50	151	8	2	16.95	28.00	15.62	0.17	0.25	0.20	0.088	-0.394	0.227
mira051enes$_\mathbf{C}$	39	19.50	153	7	1	16.95	28.00	15.62	0.17	0.25	0.20	0.093	-0.392	0.230
mira051ites$_\mathbf{C}$	36	18.00	154	10	-	16.95	26.00	9.38	0.10	0.15	0.12	0.068	-0.437	0.224
mira052ites$_\mathbf{C}$	35	17.50	154	11	-	16.95	24.00	9.38	0.10	0.15	0.12	0.071	-0.447	0.219

The virtual *combination* run was able to answer correctly 73.50% of the questions. The best performing system achieved an overall accuracy of 42% but it only gave a right answer for the 56% of the questions correctly answered by the *combination* run. Thus, we can expect improvements of the systems in a short term.
As shown in Table 10, systems generally behaved better with questions about definitions, loca-

Table 10: **Results in the tasks with Spanish as target (breakdown according to answer type)**

run	*Correct Answers*														
	Definition		*Factoid*						*Temporally restricted factoid*					*Total*	
	Or	**Pe**	**Lo**	**Me**	**Or**	**Ot**	**Pe**	**Ti**	**Lo**	**Me**	**Or**	**Ot**	**Pe**	#	%
	[25]	[25]	[21]	[17]	[22]	[19]	[20]	[19]	[6]	[7]	[6]	[6]	[7]		
inao051eses$_{\mathbf{M}}$	20	20	5	3	4	7	10	5	2	3	2	1	2	84	42.00
tova051eses$_{\mathbf{M}}$	20	20	4	3	8	4	8	7	2	3	1	1	1	82	41.00
inao052eses$_{\mathbf{M}}$	20	20	5	2	5	6	9	5	-	3	2	1	1	79	39.50
tova052eses$_{\mathbf{M}}$	20	20	4	3	8	3	3	7	2	3	1	1	2	77	38.50
upv051eses$_{\mathbf{M}}$	12	14	10	4	5	2	7	3	3	4	-	1	2	67	33.50
alia051eses$_{\mathbf{M}}$	10	10	10	3	7	3	10	2	2	3	2	2	2	66	33.00
aliv051eses$_{\mathbf{M}}$	15	8	8	3	6	2	10	5	4	1	-	-	3	65	32.50
alia052eses$_{\mathbf{M}}$	9	9	9	2	6	4	8	2	2	3	2	2	2	60	30.00
talp051eses$_{\mathbf{M}}$	16	2	11	2	5	4	8	3	1	2	2	1	1	58	29.00
talp052eses$_{\mathbf{M}}$	16	-	12	1	4	3	6	4	2	2	1	1	2	54	27.00
mira051eses$_{\mathbf{M}}$	8	9	7	3	6	4	10	1	-	2	-	-	1	51	25.50
mira052eses$_{\mathbf{M}}$	8	9	6	3	4	2	10	1	-	2	-	-	1	46	23.00
upv052eses$_{\mathbf{M}}$	-	-	10	1	2	2	8	4	3	3	-	-	3	36	18.00
combination	**23**	**24**	**19**	**10**	**16**	**10**	**16**	**10**	**5**	**5**	**3**	**2**	**4**	**147**	**73.50**
upv051enes$_{\mathbf{C}}$	6	11	7	3	3	2	5	3	3	-	-	-	2	45	22.50
mira051enes$_{\mathbf{C}}$	6	8	6	5	2	2	5	-	-	2	1	1	1	39	19.50
mira052enes$_{\mathbf{C}}$	6	8	6	5	2	2	5	-	-	2	1	1	1	39	19.50
mira051ites$_{\mathbf{C}}$	6	7	2	1	4	1	8	4	1	-	-	-	2	36	18.00
mira052ites$_{\mathbf{C}}$	5	7	2	1	4	1	8	4	1	-	-	-	2	35	17.50
combination	**11**	**16**	**10**	**5**	**7**	**5**	**9**	**6**	**3**	**2**	**1**	**1**	**2**	**78**	**39**

tions, persons and organizations. However, when the question type was measure, the accuracy tended to be lower. Indeed, this type of question has turned out to be the most difficult this year. In the factoids without temporal restrictions, the best performing system answered correctly 29.66% of the questions, a very similar accuracy comparing with the results in 2004 (see Table 12).
Concerning questions with temporal restriction, the systems with the best behavior answered correctly 34.38% of the questions, a similar result comparing with overall accuracy.

Table 11: **Results of the assessment process for questions with temporal restriction**

run	**question restriction type**		
	date	**event**	**period**
	[12]	**[10]**	**[10]**
alia051eses	3	3	5
alia052eses	3	3	5
aliv051eses	4	3	1
inao051eses	5	2	3
inao052eses	4	1	2
mira051eses	1	2	-
mira052eses	1	2	-
mira051enes	2	2	1
mira052enes	2	2	1
mira051ites	-	3	-
mira052ites	-	3	-
talp051eses	2	3	2
talp052eses	2	4	2
tova051eses	5	-	3
tova052eses	5	1	3
upv051eses	4	3	3
upv052eses	3	3	3
upv051enes	1	3	1
combination	6	8	5

As shown in Table 10, when considering the question type, the accuracy scores present small differences. Nevertheless, when the restriction type (date, event and period) is taken into account,

the differences are more important (see Table 11). It is worth mentioning that for questions restricted by event, the virtual *combination* run clearly outperforms individual systems separately (low overlapping on correct answers).
In definition questions the best performing system obtained 80% of accuracy. The improvement is remarkable considering that in the 2004 track the best systems answered correctly 70% of the questions.

Table 12: **Evolution of systems performance with Spanish as target**

	Best	**Best**	**Best**	**Best**	**Best**
Year	**Overall Acc.**	**in Fact**	**in Def**	**NIL (F)**	**r**
2003	24.5 %	24.5 %	-	0.25	-
2004	32.5 %	31.11 %	70.00 %	0.30	0.17
2005	42 %	29.66 %	80.00 %	0.38	0.56

Regarding NIL questions, the best systems achieved a recall of 0.80. F-measure improvements are also remarkable, with an increase of about 26% with respect to last year (0.30 in 2004 vs. 0.38 in 2005).
Systems have also clearly improved their confidence self-score. While in 2004 the system with higher correlation coeficience (r) reached 0.17 [2], in 2005 the highest r value was 0.56.
As shown in Table 12, the best performing systems reached and overall accuracy of 24.5%, 32.5% and 42% in 2003, 2004 and 2005, respectively (increasing +71% during the three years).

Table 13: **Results of agreement test of runs with Spanish as target language**

	# Correct	**# Correct**	**# Correct**	**# Correct**	**Disagreement**	**Kappa**	**Maximun**
run	**(Official)**	**(2nd assessor)**	**(lenient)**	**(strict)**	**#**		**variation**
es1	66	67	71	62	15	0.87	± 2 %
es2	58	63	64	57	11	0.89	± 3 %
en	45	48	51	42	11	0.87	± 4.5 %
it	35	35	39	31	10	0.86	± 2 %

In order to analyze the *interannotator agreement*, we have randomly selected 4 out of 18 runs which have been judged by two assessor with different levels of expertise. Most of the differences among assessors can be found when judging an answer as Right or as ineXact. In many cases, an assessor without experience assess as Right an answer that an experienced assessor would judge as ineXact. Table 14 shows the maximun variation of correct answers for these four runs (average = ± 2.9%).

Table 14: **Results of agreement test of runs with Spanish as target language**

	# Correct	**# Correct**	**# Correct**	**# Correct**	**Disagreement**	**Kappa**	**Maximun**
run	**(Official)**	**(2nd assessor)**	**(lenient)**	**(strict)**	**#**		**variation**
es1	66	67	71	62	15	0.87	± 2 %
es2	58	63	64	57	11	0.89	± 3 %
en	45	48	51	42	11	0.87	± 4.5 %
it	35	35	39	31	10	0.86	± 2 %

Finally we can conclude that both the improvement in systems' self-evaluation, the scores obtained by the participanting systems (73.50% in combination, 42% individually), and the systems' evolution during the last three years, let us expect a significant improvement in Spanish question answering technologies in the near future.

Table 15: **Results in the tasks with Finnish as target**

run	Right		W	X	U	Right			NIL [20]			CWS	K1	r
	#	%	#	#	#	% F [111]	% D [60]	% T [29]	P	R	F			
hels051fifi$_M$	46	23.00	131	23	-	18.92	25	34.98	0.13	0.35	0.19	0.090	-0.202	0.064
hels052fifi$_M$	38	19.00	140	22	-	15.32	23.33	24.14	0.12	0.30	0.17	0.074	-0.230	0.093

4.5 Finnish as Target

The year 2005 was the first year when Finnish existed as a target language. Only one group submitted runs for this task, and both of the runs were monolingual. The artificial combination run presented in table 16 shows that the upper bound on the performance of a system that would merge the results of the existing runs and somehow select the right answers from the combined pool of candidate answers is 26.50%. This is by far the lowest monolingual combination run score among the participating languages. The next one is bulgarian with a combination score of 36.00 % (see Table 4). However, when we calculate the average score for the monolingual runs of each target language, we can see that Finnish is not very far behind, for the average accuracy of the Finnish runs is 21.00%, that of the Bulgarian ones is 23.00%, that of the Italian ones is 24,08%, that of the French ones is 25,20%, and so on. The confidence scores that the systems having Finnish as target assign to the answers only very faintly reflect the assessor's opinion on the correctness of the answer, as can be seen from the correlation coefficient between the system's score and correctness (r) in Table 15.

Table 16: **Results in the tasks with Finnish as target (breakdown according to answer type)**

run	*Correct Answers*														
	Definition		*Factoid*						*Temporally restricted factoid*					*Total*	
	Or [27]	Pe [33]	Lo [21]	Me [10]	Or [15]	Ot [20]	Pe [28]	Ti [17]	Lo [4]	Me [5]	Or [5]	Ot [5]	Pe [10]	#	%
hels051fifi$_M$	6	9	4	1	2	1	9	4	1	4	-	1	4	46	23.00
hels052fifi$_M$	8	6	4	1	2	-	7	3	1	3	-	1	2	38	19.00
combination	**8**	**10**	**5**	**1**	**3**	**1**	**11**	**4**	**1**	**4**	**-**	**1**	**4**	**53**	**26.50**

The evaluation of the Finnish answers was not straightforward because the evaluation guidelines [1] don't discuss word affixes with regard to the exactness of the answers. Finnish is a highly inflecting language where each noun, for example, has 15 different cases. In addition to cases, nouns can also contain possessive suffixes and clitics. Most of the answers to the CLEF questions are noun phrases. The cases, possessive suffixes and clitics typically express meanings that are in the other target languages of the evaluation campaign expressed by separate words such as prepositions, pronouns and adverbs. Thus, one single word in Finnish may convey consederably more information than a single word in the other target languages. For example, the word *talossanikin* means *also in my house.* Our understanding of the guidelines was that the answer should be taken from text as such, without any modifications, such as lemmatization. Now, due to the rich affixing, the answer that is not lemmatized may contain additional information that disturbs the evaluator, and he is tempted to judge the answer inexact. However, judging as inexact all those answers that are not in the form required by the question could not be done, because that is not required according to the guidelines. When deciding how to assess the Finnish answers, we observed how the judgements had been done with regard to cases in the other target languages. For example, in German, the case may cause modifications in the determiner. However, those answers whose head noun is not in the nominative case even though that is the case requested by the question, are marked as correct. For example: Question: *62 D PER Wer ist Goodwill Zwelithini?* Answer:

Table 17: **Results in the tasks with French as target**

run	**Right** #	%	**W** #	**X** #	**U** #	**Right** % **F** [120]	% **D** [50]	% **T** [30]	**NIL** [20] **P**	**R**	**F**	**CWS**	**K1**	**r**
syna051frfr$_{\mathbf{M}}$	128	64.00	62	8	2	59.17	86.00	46.67	0.23	0.25	0.24	-	-	-
tova052frfr$_{\mathbf{M}}$	70	35.00	120	10	-	27.50	66.00	13.33	0.14	0.30	0.19	-	-	-
tova051frfr$_{\mathbf{M}}$	69	34.50	121	10	-	26.67	66.00	13.33	0.13	0.25	0.17	-	-	-
upv051frfr$_{\mathbf{M}}$	46	23.00	143	7	4	17.50	46.00	6.67	0.06	0.10	0.07	0.115	-0.048	0.210
hels051frfr$_{\mathbf{M}}$	35	17.50	156	8	1	16.67	22.00	13.33	0.10	0.45	0.17	0.108	-0.196	0.281
upv052frfr$_{\mathbf{M}}$	34	17.00	160	5	1	15.00	20.00	20.00	0.07	0.20	0.10	0.073	-0.057	0.207
lire051frfr$_{\mathbf{M}}$	33	16.50	145	20	2	15.83	14.00	23.33	0.09	-	0.09	0.072	-0.358	0.260
hels052frfr$_{\mathbf{M}}$	33	16.50	157	10	-	15.00	22.00	13.33	0.09	0.40	0.15	0.097	-0.230	0.247
lina051frfr$_{\mathbf{M}}$ *	29	14.50	144	21	3	17.95	6.00	16.67	0.15	0.20	0.17	0.048	-0.470	0.151
lcea051frfr$_{\mathbf{M}}$	28	14.00	165	3	4	18.33	0.00	20.00	0.33	0.05	0.09	-	-	-
syna051enfr$_{\mathbf{C}}$	79	39.50	108	10	3	30.25	72.00	22.58	0.14	0.30	0.19	-	-	-
syna051ptfr$_{\mathbf{C}}$	73	36.50	115	9	3	26.67	68.00	23.33	0.07	0.15	0.10	-	-	-
syna051itfr$_{\mathbf{C}}$	51	25.50	136	11	2	15.00	54.00	20.00	0.13	0.45	0.21	-	-	-

* Results calculated over 197 questions.

R 0062 dem König der Zulus[2]. Thus, we decided to judge as correct in Finnish also those answers that are not in the form required by the question. For example: Question: *65 F PER Kuka on ohjannut elokuvan Hamlet liikemaailmassa?* Answer: *R 0067 Mika Kaurismäen*[3]. In fact, most of the problematic question forms in the test set for Finnish are of the type where the answer is given in the genetive case and the case required by he question is the nominative case.

4.6 French as Target

Seven research groups took part in evaluation tasks using French as target language: Synapse Développement (France), CEA-LIST/LIC2M (France), LIMSI-LIR (France), Université de Nantes, LINA (France), Helsinki University (Finland), Universitat Politècnica de València, UPV (Spain) and TOVA, a joint system between UPV and the Instituto Nacional de Astrofísica Óptica y Electrónica (Mexico). All participating groups took part in the monolingual task: four groups submitted one run and three groups submitted two runs FR-FR. Only Synapse Développement took part in the bilingual tasks. This group submitted three runs, one run per source language: Italian, English and Portuguese. Table 17 shows the results of the assessment of the thirteen submitted runs. This year, many groups participated in the Question Answering tasks with French as a target. It appears that the number of participants for the French task has increased significantly: seven this year as opposed to one last year. The best results were obtained by Synapse Développement for one of the monolingual runs (syna051frfr). This group ranked 2nd and 3rd in the two English-French and Portuguese-French runs which is better than all the other monolingual French runs. The two monolingual runs by the spanish TOVA group reached the 4th and 5th positions.

The correct answers given for all the runs are presented in table 18, sorted by type of answer (location, measure, organization, etc.). The results show the limits of the system developed by Synapse Développement, which obviously lie in factoid-other (9/20), factoid-measure (10/20) and factoid-time (11/20), whereas results are much better for definition and factoid-person questions. The aim of the virtual run called combination is to provide an upperbound on the possible performance of a system that would merge the existing runs and somehow select the right answers from the combined pool of candidate answers. The best run (syna051frfr) is able to supply 76.19% of the correct answers of combination. This ratio could be enhanced if results for factoid-measure or factoid-time questions were better.

[2] The question requires the head noun of the answer to be in the nominative case - *der König* - instead of the dative case - *dem König.*

[3] The question requires the answer to be in the nominative case - *Mika Kaurismäki* - instead of the genetive case - *Mika Kaurismäen.*

Table 18: **Results in the tasks with French as target (breakdown according to answer type)**

run	*Correct Answers*														
	Definition		*Factoid*						*Temporally restricted factoid*					*Total*	
	Or [25]	**Pe** [25]	**Lo** [20]	**Me** [20]	**Or** [20]	**Ot** [20]	**Pe** [20]	**Ti** [20]	**Lo** [6]	**Me** [2]	**Or** [5]	**Ot** [10]	**Pe** [7]	#	%
syna051frfr$_{\mathbf{M}}$	21	22	12	10	13	9	16	11	4	1	2	4	3	128	64.00
tova052frfr$_{\mathbf{M}}$	19	14	3	3	7	5	8	7	1	-	1	1	1	70	35.00
tova051frfr$_{\mathbf{M}}$	19	14	3	2	8	5	7	7	1	-	1	1	1	69	34.50
upv051frfr$_{\mathbf{M}}$	9	14	3	5	3	2	4	4	1	-	-	-	1	46	23.00
hels051frfr$_{\mathbf{M}}$	4	7	5	5	3	1	3	3	-	1	1	2	-	35	17.50
upv052frfr$_{\mathbf{M}}$	-	10	3	2	2	1	4	6	3	-	1	1	1	34	17.00
lire051frfr$_{\mathbf{M}}$	3	4	1	4	5	1	2	6	3	1	1	1	1	33	16.50
hels052frfr$_{\mathbf{M}}$	4	7	5	3	2	1	3	4	-	1	1	2	-	33	16.50
lina051frfr$_{\mathbf{M}}$ *	1	2	3	2	2	3	6	5	1	-	1	1	2	29	14.50
lcea051frfr$_{\mathbf{M}}$	-	-	3	4	5	-	6	4	1	-	1	-	4	28	14.00
combination	**23**	**23**	**15**	**16**	**16**	**14**	**18**	**16**	**5**	**2**	**3**	**5**	**5**	**161**	**80.5**
syna051enfr$_{\mathbf{C}}$	21	15	8	6	3	9	6	4	2	-	1	3	1	79	39.50
syna051ptfr$_{\mathbf{C}}$	17	17	6	4	8	7	4	3	4	-	2	1	-	73	36.50
syna051itfr$_{\mathbf{C}}$	15	12	4	6	2	4	1	1	-	1	1	3	1	51	25.50
combination	**21**	**20**	**10**	**10**	**8**	**10**	**8**	**6**	**5**	**1**	**2**	**3**	**2**	**106**	**53**

* Results calculated over 197 questions.

The main problem encountered during the assessment of answers was related to the temporally restricted factoid questions. This year and for the first time in CLEF this kind of questions was included in the test sets. We thought that the generation of this kind of questions would be relatively easy, but did not foresee that the assessment on those questions would be so difficult.

In fact, many temporally restricted factoid questions have not been built properly as there was no logic of restriction at all. The question ”In which famous capital was the Eiffel Tower built in 1889?” is a good example. Here, ”in 1889” is a redundant information rather than a temporally restriction and will be ignored by the system: the correct answer returned with a document associating the Eiffel Tower to Paris will be a right answer even if it does not specify that the Eiffel Tower was built in 1889.

Therefore, from the beginning of the assessment phase on, many questions arise such as ”Should the date be included in the document joined to the answer?”, ”Should all the items included in the question be found in the document in order to consider the answer as correct?”. Now we know how to handle those temporally restricted factoid questions and such problems should not occur next year.

This year, as far as French language is concerned, the best system obtained very good results : 128 correct answers out of 200. In all the QA@CLEF tracks, these are the best results ever obtained for the French used as target language. Moreover, we could see a growing interest in Question Answering from the European research community: the QA@CLEF-2005 attracted more participants in evaluation tasks using French as target language than the previous editions. In addition, the benchmark resources built for these evaluations contributed to the development and the improvment of systems, and could be used again as training resources in the next edition.

4.7 Italian as Target

Three groups participated in the Italian monolingual task, and no one in the other bilingual tasks with Italian as target. A total of six runs were submitted, two each research group: ITC-Irst, the Universidad Politécnica de Valencia (UPV) and a joint experiment by UPV and the Mexican INAOE (Instituto Nacional de Astrofísica, Óptica y Electrónica). As table 19 shows the best system (the one developed by UPV and INAOE) answered correctly to 27.5% of the questions, and the other two systems achieved similar results.

In 2004 two teams had participated in the Italian monolingual task, submitting a total of 3 runs. The best performer had an overall accuracy of 28%, while the average performance was

Table 19: **Results in the tasks with Italian as target**

run	**Right**		**W**	**X**	**U**	**Right**			**NIL** [20]			**CWS**	**K1**	**r**
	#	%	#	#	#	**% F** [120]	**% D** [50]	**% T** [30]	**P**	**R**	**F**			
tova052itit$_{\mathrm{M}}$	55	27.50	135	10	-	23.33	42.00	20.00	0.15	0.55	0.24	-	-	-
tova051itit$_{\mathrm{M}}$	53	26.50	138	9	-	21.67	42.00	20.00	0.16	0.55	0.24	-	-	-
upv_051itit$_{\mathrm{M}}$	51	25.50	142	6	1	20.00	44.00	16.67	0.10	0.15	0.12	0.156	0.012	0.316
upv_052itit$_{\mathrm{M}}$	48	24.00	148	4	-	15.83	50.00	13.33	0.06	0.15	0.09	0.125	-0.200	0.202
irst051itit$_{\mathrm{M}}$	44	22.00	137	17	2	19.17	38.00	6.67	0.17	0.20	0.18	0.129	-0.197	0.267
irst052itit$_{\mathrm{M}}$	38	19.00	145	14	3	14.17	38.00	6.67	0.40	0.10	0.16	0.100	-0.301	0.071

25.1%. In 2005 the task itself attracted more research groups, and though the best system was approximately as good as the one of last year, the average overall accuaracy is slightly worse (i.e. 24%), which probably means that the Italian monolingual test set was more challenging in 2005. As far as the types of questions are concerned, it is interesting to notice that definitional questions proved to be easier than factoids. Between 38 and 50% of definitional got a correct answer, while temporally restricted questions were tougher for the three participating systems. Eleven questions (no. 3, 20, 30, 60, 65, 84, 85, 107, 113, 116 and 124) received a correct answer in all the six submitted runs, and five among them are definition questions referred to a person. This suggests that this type of questions have often a straightforward answer that appears between brackets or in appositive form within the text. Table 20 shows that the factoids with *location*, *person* and *time* as answer type were the easiest for systems, and if the three systems had worked together, they could have achieved an overall accuracy of 46.5%, which encourages research groups to share tools and resources in the future.

Table 20: **Results in the tasks with Italian as target (breakdown according to answer type)**

run	***Correct Answers***														
	Definition		***Factoid***						***Temporally restricted factoid***					***Total***	
	Or [25]	**Pe** [25]	**Lo** [19]	**Me** [21]	**Or** [21]	**Ot** [19]	**Pe** [20]	**Ti** [20]	**Lo** [4]	**Me** [4]	**Or** [3]	**Ot** [8]	**Pe** [11]	#	%
tova052itit$_{\mathrm{M}}$	11	10	4	1	7	1	8	7	1	1	-	1	3	55	27.50
tova051itit$_{\mathrm{M}}$	11	10	4	1	7	1	6	7	1	1	-	1	3	53	26.50
upv_051itit$_{\mathrm{M}}$	10	12	6	2	3	2	7	4	1	1	-	-	3	51	25.50
upv_052itit$_{\mathrm{M}}$	11	14	5	-	3	1	6	4	1	1	-	-	2	48	24.00
irst051itit$_{\mathrm{M}}$	5	14	7	-	4	2	5	5	1	-	-	-	1	44	22.00
irst052itit$_{\mathrm{M}}$	5	14	3	-	4	1	3	6	1	-	-	-	1	38	19.00
combination	**17**	**20**	**9**	**3**	**8**	**4**	**10**	**11**	**3**	**2**	**-**	**1**	**5**	**93**	**46.5**

The manual assessment procedure was the same as it was in 2004. Two assessors had a brief training session (based on the 2004 submissions) that aimed at making them familiar with the evaluation tool interface and at solving preliminary doubts. Both assessors judged all the six runs and then the answers with different judgments were double-checked and received a third, final judgment. Table 21 gives the number of different judgments per run and the inter-assessor kappa coefficient, which is quite high (average value is 0.874).

Table 21: **Inter-assessor agreement in the evaluation of the Italian runs**

	disagreement	
run	different judgments (#)	kappa coefficient
tova052itit	10	0.895
tova051itit	11	0.882
upv_051itit	8	0.909
upv_052itit	9	0.895
irst051itit	17	0.828
irst052itit	15	0.839

A total of 70 disagreement cases were registered, most of them involved the judgment couples R-X (11 cases), R-W (13 cases), U-W (10 cases) and above all X-W (31 cases). Clearly, the evaluation guidelines did not deal extensively with answer exactness, so assessors had some difficulties in deciding which portion of an answer-string was acceptable and which not. In most of the cases (i.e. 26) where an assessor assigned X and the other W, the third and final judgment was W.

4.8 Dutch as Target

This year two teams that took part in the QA@CLEF track used Dutch as their target language: the University of Amsterdam and the University of Groningen. In total, three runs were submitted, all using Dutch as the source language. All runs were assessed by two assessors, with very inter-assessor agreement (0.950 for gron051nlnl, and 0.976 for uams051nlnl and uams052nlnl). The results of the evaluation for all runs are provided in Tables 22 and 23.

Table 22: **Results in the tasks with Dutch as target**

run	**Right**		**W**	**X**	**U**	**Right**			**NIL** [20]			**CWS**	**K1**	**r**
	#	**%**	**#**	**#**	**#**	**% F** [114]	**% D** [60]	**% T** [26]	**P**	**R**	**F**			
gron051nlnl$_{\mathrm{M}}$	99	49.50	79	18	4	54.39	50.00	26.92	0.46	0.30	0.36	0.382	0.071	0.302
uams051nlnl$_{\mathrm{M}}$	88	44.00	79	28	5	47.37	45.00	26.92	0.77	0.50	0.61	-	-	-
uams052nlnl$_{\mathrm{M}}$	88	44.00	78	29	5	48.25	43.33	26.92	0.77	0.50	0.61	-	-	-

When scored in terms of the percentage of correct (i.e, correct and exact and supported) answers, the run labeled gron051nlnl (submitted by the University of Groningen) clearly outperforms the two runs submitted by the University of Amsterdam: 49.50% vs. 44% and 44%. When compared to the correct answers in the Groningen run, many of the inexact answers in the Amsterdam runs are caused by incorrect definitions; here's an example.

0094 NLNL Wat is Eyal?
gron051nlnl: militante joodse groep
uams051nlnl: leider van de extreem-rechtse groep

This observation is confirmed if we take a closer look. In the 200 questions, six initial words occur more than ten times: *Wie* (*Who*), *Wat* (*What*), *Hoe* (*How*), *Welke* (*Which*), *Waar* (*Where*) and *In* (*In*). The performance of the questions with four of the six initial words is similar for the three runs. For *Wat*, Groningen obtains 67% right and Amsterdam 39%. This difference is mainly caused by the problem with the definition answers just mentioned. For *Hoe*, Groningen obtains 63% and Amsterdam 36%. Seven of the eight *Hoe* questions for which only Groningen found the answer, were of the format *Hoe heet DEFINITION?* (*What is the name of DEFINITION?*).

All in all, the Groningen run performs noticeably better than the Amsterdam runs in terms of *precision*—this is clear from the differences in answers labeled X (inexact): only 18 for Groningen, and as many as 28 and 29 for Amsterdam.

If we drill down a bit further, and consider the detailed results in Table 23, we see that Groningen outperforms Amsterdam on Organisations in the Definitions category, and on Other questions in the Factoid category; Amsterdam is slightly better in Person definitions. On other categories, the differences are very minor or non-existent. There is, however, a noticeable difference in performance on NIL questions, with Amsterdam achieving far higher F-scores than Groningen.

To conclude, let's adopt a somewhat alternative perspective. The differences between the Groningen run and the Amsterdam are mainly in the number of inexact answers; in terms of the number of unsupported or wrong answers the differences are negligible. Put differently, in terms of the number of answers that are "helpful" [4], i.e., that would help a user meet her information needs, the three runs all perform at the same level: 117 helpful (i.e., correct or inexact) for the Groningen run, and 116 and 117 helpful for the two Amsterdam runs.

Table 23: **Results in the tasks with Dutch as target (breakdown according to answer type)**

run	*Correct Answers*														
	Definition		*Factoid*						*Temporally restricted factoid*					*Total*	
	Or [24]	**Pe** [36]	**Lo** [30]	**Me** [9]	**Or** [11]	**Ot** [20]	**Pe** [35]	**Ti** [9]	**Lo** [5]	**Me** [4]	**Or** [1]	**Ot** [4]	**Pe** [12]	#	%
gron051nlnl$_{\mathbf{M}}$	16	14	13	4	3	12	25	5	3	1	-	-	3	99	49.50
uams051nlnl$_{\mathbf{M}}$	9	18	14	2	3	7	23	5	2	-	-	-	5	88	44.00
uams052nlnl$_{\mathbf{M}}$	8	18	14	2	4	7	23	5	2	-	-	-	5	88	44.00
combination	**17**	**24**	**21**	**4**	**6**	**13**	**31**	**7**	**5**	**1**	**-**	**-**	**7**	**136**	**68.00**

4.9 Portuguese as Target

In 2005 there were five runs with Portuguese as target, submitted by three different research teams: In addition to the two participants from last year, SINTEF with the Esfinge system and the University of Évora, we had a newcomer from industry, Priberam, a Portuguese company specialized in NLP products. Although a collection of Brazilian Portuguese news was added to the CLEF collection, no Brazilian participants turned up as yet for CLEF.

Table 24: **Results in the tasks with Portuguese as target**

run	**Right**		**W**	**X**	**U**	**Right**			**NIL** [18]			**CWS**	**K1**	**r**
	#	%	#	#	#	**% F** [135]	**% D** [42]	**% T** [23]	**P**	**R**	**F**			
prib051ptpt$_{\mathbf{M}}$	129	64.50	55	13	3	67.41	64.29	47.83	0.50	0.11	0.18	-	-	-
ptue051ptpt$_{\mathbf{M}}$	50	25.00	125	22	3	21.48	35.71	26.09	0.10	0.67	0.18	0.250	-0.500	0.000
esfg051ptpt$_{\mathbf{M}}$	46	23.00	139	11	4	23.70	16.67	30.43	0.21	0.78	0.33	-	-	-
esfg052ptpt$_{\mathbf{M}}$	43	21.50	145	10	2	23.70	14.29	21.74	0.22	0.78	0.34	-	-	-
esfg051enpt$_{\mathbf{C}}$	24	12.00	165	9	2	11.11	14.29	13.04	0.12	0.78	0.20	-	-	-

Table 24 presents the five runs. This year there was a first crosslingual run, from English to Portuguese, by Esfinge, with significantly worse results than the monolingual runs, as might be expected. As to the monolingual results, the Esfinge system showed some improvement as compared to last year, although its best run was still unable to equal PTUE system's score. PTUE's results, however, were slightly worse than last year's. The clear winner in all respects was Priberam's system, which, in fact, was the best participating system in the whole QA@CLEF. Table 25 breaks down the correct answers by kind of entity, as well as provides a combination score: a question is considered answered if any system has been able to provide a right answer (assuming that a user would be able to check easily, in case of multiple answers, the right one). In this, we see that Portuguese language ranks as second, after French.

Table 25: **Results in the tasks with Portuguese as target (breakdown according to answer type)**

run	*Correct Answers*														
	Definition		*Factoid*						*Temporally restricted factoid*					*Total*	
	Or [15]	**Pe** [27]	**Lo** [30]	**Me** [17]	**Or** [21]	**Ot** [15]	**Pe** [37]	**Ti** [15]	**Lo** [5]	**Me** [1]	**Or** [2]	**Ot** [6]	**Pe** [9]	#	%
prib051ptpt$_{\mathbf{M}}$	12	15	26	11	7	8	25	14	2	-	-	4	5	129	64.50
ptue051ptpt$_{\mathbf{M}}$	5	10	10	1	3	3	10	2	1	-	-	2	3	50	25.00
esfg051ptpt$_{\mathbf{M}}$	1	6	9	4	5	2	9	3	1	-	1	3	2	46	23.00
esfg052ptpt$_{\mathbf{M}}$	-	6	8	3	3	-	13	5	-	-	1	2	2	43	21.50
combination	**12**	**22**	**28**	**13**	**11**	**10**	**27**	**15**	**3**	**-**	**1**	**6**	**7**	**155**	**77.50**
esfg051enpt$_{\mathbf{C}}$	-	6	3	1	3	1	5	2	-	-	1	2	-	24	12.00
combination	**-**	**6**	**3**	**1**	**3**	**1**	**5**	**2**	**-**	**-**	**1**	**2**	**-**	**24**	**12.00**

Another relevant remark is that definitions do not seem to be more difficult on average than factoid questions, as was the case last year. We believe, however, that this is due to a considerable simplification of precisely what "definition questions" are, where they boil down to mainly ask for a person's profession or title. We did some further analysis of the results in order to have other measures of confidence in the systems, which are displayed in table 26. We looked specifically at (i) the cases where no answer was given (*null answer*), which keep the user in a state of ignorance, no matter the system was right in providing the null answer or wrong because it could not find it; (ii) the cases where any user could at once see the answer was rubbish (*rubbish*); and (iii) the cases where the wrong answers could be misleading (*dangerous*). Of course it depends on the ignorance of the questioner, and we were very conservative in imagining total ignorance. Probably most of the "dangerous" questions would at once be spotted as system's mistakes by an ordinary user – or at least arise some suspicion.

Table 26: **Results in the tasks with Portuguese as target (breakdown of bad answers)**

	Incorrect or null answers		
run	null answer	rubbish	dangerous
prib051ptpt	4	13	43
ptue051ptpt	117	3	20
esfg051ptpt	68	41	49
esfg052ptpt	65	34	63
esfg051enpt	121	21	40

The results show that the PTUE system is both the most reliable (less non-NIL wrong answers) and the most conservative system (most empty answers), the more "dangerous" one being Esfinge.

5 Conclusions

This paper presented the multilingual Question Answering evaluation campaign organized at CLEF 2005. QA@CLEF considerably increased both in number of participants -we are now closer to the Question Answering track at TREC- and also in the number of languages involved. It is also relevant that this year we were able to activate a task with Bulgarian as a target, a language of a new EU member country. A pilot cross-language task with Indonesian as source and English as target has been also activated.

Being the organization of the task at its third year, is now well tested, although involving nine different Institutions of as many different countries, and showed to be able to support the high number of exchanges required by the organization of the task. This is particularly significative considering that all the organizations involved in QA@CLEF guarantee their support on a voluntary basis.

The increased number of participants allowed to carried out a number of interesting comparisons among systems participating at the same task (this was one of the drawback of the 2004 campaign). In addition, it is worth to mention that Question Answering techniques for European languages, being mainly based on NLP tools and resources for the respective languages, demand for better tool and resources. In a cross-language perspective the integration of such resources is also crucial.

Finally, having (at least partially) achieved its goal to promote Question Answering for European languages, there is now a quite large scientific community in Europe on Question Answering, QA@CLEF is now ready to propose its own view on QA, designing a roadmap for the next years multilingual QA systems.

Acknowledgements

Anselmo Peñas and Valentín Sama Rojo have been partially supported by the Spanish Government under project R2D2-Syembra TIC-2003-07158-C04-02.

Maarten de Rijke was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 612-13-001, 612.000.106, 612.000.207, 612.066.302, 612.069.006, and 640.001.501.

References

[1] The CLEF QA Track coordinators. QA@CLEF 2005 Guidelines, 2005. http://clef-qa.itc.it/2005/guidelines.html.

[2] Jesús Herrera, Anselmo Peñas, and Felisa Verdejo. Question answering pilot task at clef 2004. *Proceedings of CLEF 2004. Lecture Notes in Computer Science. Springer-Verlag*, (3491):581–590, 2005.

[3] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. In F. Borri In C. Peters, editor, *Results of the CLEF 2004 Cross-Language System Evaluation Compaign*, Bath, U.K., 2004. Working Notes for the CLEF 2004 Workshop.

[4] K. Spark Jones. Is question answering a rational task? In R. Bernardi and M. Moortgat, editors, *Questions and Answers: Theoretical and Applied Perspectives (Second CoLogNET-ElsNET Symposium)*, pages 24–35, 2003.