# English run of Synapse Développement at Entrance Exams 2014

Dominique Laurent, Baptiste Chardon, Sophie Nègre, Patrick Séguéla

Synapse Développement, 5 rue du Moulin-Bayard, 31000 Toulouse
{dlaurent, baptiste.chardon, sophie.negre,
patrick.seguela}@synapse-fr.com

**Abstract.** This article presents the participation of Synapse Développement to the CLEF 2014 Entrance Exam campaign (QA track). Since fifteen years, our company works on Question Answering domain. Recently our work concentrated on Machine Reading and Natural Language understanding. Thus, the Entrance Exam evaluation was an excellent opportunity to measure the results of this work. The developed system is based on a deep syntactic and semantic analysis with anaphora resolution. The results of this analysis are saved in sophisticated structures based on clause description (CDS = Clause Description Structure). For this evaluation, we added a dedicated module to compare CDS from texts, questions and answers. This module measures the degree of correspondence between these elements, taking into account the type of question, which means the type of answer awaited. We participate in English and French languages; this article focuses on the English run, comparing it with the French run whose final results were better. However our run for English obtains the best results in this language.

**Keywords**: Question Answering, Machine Reading, Natural Language Understanding.

## 1 Introduction

The Entrance Exams evaluation campaign uses real reading comprehension texts coming from Japanese University Entrance Exams (the Entrance Exams corpus for the evaluation is delivered by NII's Todai Robot Project [13] and NTCIR RITE). These texts are intended to be used to test the level of English of future students and represent an important part in Japanese University Entrance Exams[1]. As claimed by the organizers of this campaign: " *The challenge of "Entrance Exams" aims at evalu-*

[1] See in References [3] and [6] but also http://www.ritsumei.ac.jp/acd/re/k-rsc/lcs/kiyou/4-5/RitsIILCS_4.5pp.97-116Peaty.pdf

*ating systems under the same conditions humans are evaluated to enter the University"*[2].

Our Machine Reading system is based on a major hypothesis: The text, in its structure and in its explicit and implied syntactic functions, contains enough information to allow Natural Language Understanding with a good accuracy. So our system does not use any external resources, i.e. Wikipedia, DbPedia and so on. Our system uses only our linguistic modules (parsing, word sense disambiguation, named entities detection and resolution, anaphora resolution) and our linguistic resources (grammatical and semantic information on more than 300,000 words and phrases, global taxonomy on all these words, thesaurus, families of words, converse relation dictionary (for example, "sell" and "buy", or "marry"), and so on). These software modules and linguistic resources are the results of more than twenty years of development and are considered and evaluated as the state of art for French and English.

Our Machine Reading system and the Multiple-Choice Question-Answering system needed for Entrance Exams use a database built with the results of our analysis that results in a set of Clause Description Structures (CDS) to be described in the second chapter of this article.

The Entrance Exams corpus was composed this year of 12 texts with a total of 56 questions. Knowing that for each question 4 answers are proposed, the total number of choices/options was 224. Organizers of the evaluation campaign allow the systems to leave some questions unanswered if theses systems are not confident in the correctness of the answer. We did not use this opportunity but we will give in chapter 3 some results when leaving unanswered questions where the probability of the best answer is too low and other results when leaving unanswered questions where the probability of the best answer is not superior or equal to the double of the probability of the second best answer.

## 2 Machine Reading System architecture

For Entrance Exams, similar treatments are made for texts, questions and answers but the results of these treatments are saved in three different databases, allowing the final module to compare the Clause Description Structures (CDS) from text and answers to measure the probability of correspondence between CDS from text and CDS from answers. . Figure 1 shows the the global architecture of our system.

[2] http://nlp.uned.es/entrance-exams/

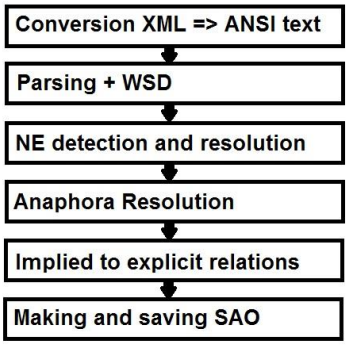


Figure 1. Description of the system

### 2.1 Conversion from XML into text format

The XML format allows our system to distinguish text, questions and answers. So, it's very useful but our different linguistic modules manage only text format. So the first operation is to extract text, then each question and the corresponding answers in text format.

### 2.2 Parsing, Word Sense Disambiguation, Named Entities detection

We use our internal parser which begins by a lexical disambiguation (is it a verb? a noun? a preposition? and so on) and a lemmatization. Then the parser splits the different clauses, groups the phrases, sets the part of speech and searches all grammatical functions (subject, verb, object, direct or indirect, other complements).

Then, for all polysemous words, a Word Sense Disambiguation module detects the sense of the word. For English, this detection is successful in 82% of word senses (87% for French with a higher number of polysemous words and a higher number of senses for each word). The senses disambiguated are directly linked in our internal taxonomy.

A named entity detector groups the named entities. The Named Entities detected are : names of persons, organizations and locations, but also functions (director, student, etc.), time (relative or absolute), numbers, etc. These entities are linked between them when they refer to the same entity (for example "Dominique Strauss-Kahn" or "DSK", "Toulouse" or "la Ville rose", etc). This module is not very useful for this Entrance Exams campaign but for time entities.

### 2.3 Anaphora resolution

In English, we consider as anaphora all the personal pronouns (I, me, he, him, she, her, it, we, us, you, they, myself, yourself, himself, herself, itself, ourselves, yourselves, themselves), all demonstrative pronouns and adjectives (this, that, these, those), all possessive pronouns and adjectives (my, mine, his, her, its, our, ours, your,

yours, their, theirs) and, of course, the relative pronouns (who, whom, whose, which, what, that) and the pronouns "one" and "ones".

During the parsing, the system builds a table with all possible referents for anaphora (proper nouns, common nouns, phrases, clauses, citations) with a lot of grammatical and semantic information (gender, number, type of named entity, category in the taxonomy, sentence where the referent is located, number of references for this referent, etc.) and, after the syntactic parsing and the word sense disambiguation, we resolve the different anaphora in the sentence by comparison with our table of referents. Our results at this step are slightly inferior to the state of the art, especially for demonstrative adjectives and for relative pronouns. Some errors come, of course, of errors in lexical disambiguation, for example confusion between personal pronoun and possessive adjective (his, her, for example in "*At first Mrs. Tortino thought he would offer her money for her hom*e", the parser considers "her" as a possessive adjective linked to "money") or demonstrative pronoun and relative pronoun (that).

### 2.4 Implied to explicit relations

When there are coordinate subjects or objects (for example "Dad and Mom"), our system keeps the trace of this coordination. For example with the coordination "Dad and Mom" the system will save three different CDS, one with the coordinate subject and two for each term of the subject. The aim of this division is to find possible answers with only one term of the coordination. But, beyond this very simple decomposition, our analyzer operates more complex operations. For example, in the sentence "*Certainly, many animals, especially the young, engage in behavior that seems like play*", extracted from third text of this evaluation, our system will add "animals" after "the young", this type of completion is very close of anaphora resolution but different because the system tries to add implied information, which are generally nouns or verbs. This mechanism exists also for the CDS structures as described in the next paragraph.

### 2.5 Making and saving CDS

We describe in this Section the main features of CDS structures. First we consider the attribute as an object (that could be discussed, but it allows one model of structure only). The main components of the structure are descriptions of a clause, normally compound of a subject, a verb and an object or attribute. Of course the structure allows many other components, for example indirect object, temporal context, spatial context... Each component is a sub-structure with the complete words, the lemma, the possible complements, the preposition if any, the attributes (adjectives) and so on.

For verbs, if there is some modal verb, only the last verb is considered but the modality relation is kept in the structure. Of course negation or semi-negation (forget to) are also attributes of the verb in the structure. If a passive form is encountered, the real subject becomes the subject of the CDS and the grammatical subject becomes the object. When the system encountered possessive adjective, a specific CDS is created with a link of possession. For example, in the sentence "*He often talked to me about*

*his home in Wisconsin*" where "*he*" is the referent of a Winnebago Indian, the system creates one CDS with "*Winnebago Indian*" as subject, "*talk*" as verb, "*I*" as indirect objet and "*home*" as direct object. But the system creates also another CDS with "*Winnebago Indian*" as subject, "*have*" as verb (possession), "*home*" as object and "*Wisconsin*" as spatial context.

New CDS are also created when there is a converse relation. For example, in the sentence ""*Don't worry about it, Dad," Patrick said*.", where "Dad" is the author (anaphora resolution from precedent sentences), the system will extract one CDS with "*I*" (the author) as subject, "*be*" as verb, "*father*" as object and "*Patrick*" as complement of "*father*", but also another CDS with "*Patrick*" as subject, "*be*" as verb, "*son*" as object and "*I*" as complement of "*son*". The system manages 347 different converse relations, for example the classical "sell" and "buy", or "mary", or "manager" and "employee", but also geographic terms (south/north, under/on top...) and time terms (before/after, previous/next...). For all these links, two CDS are created.

Links between CDS are also saved. For example, in the sentence "*He felt that she looked just as he had imagined*", we have three CDS ("He felt", "she looked" and "he has imagined") but the object of SAO1 is SAO2 and the object of SAO3 is also SAO2. Other relations like "aim", "cause", "consequence", "judgment", "opinion" and so on are also saved and are important when the system matches the CDS of the text and the CDS of the possible answers. At the end, after all these extensions, we can consider that a real semantic role labelling is performed.

Finally the system saves also "referents", which are proper and common nouns found in the sentences, after anaphora resolution. These referents are especially useful when the system do not find any correspondence between CDS, knowing that the frequencies in text and in usual vocabulary are arguments of the referent structures.

A specific difficulty of Entrance Exams corpus is that it is frequently spoken language with dialogs like in novels. It needs a deep analysis of the characters as you can imagine with some sentences like " *"I don't want to go to a new school. I like my school here. And what about my friends?" "Don't worry, Elena. You'll make new friends." I didn't want new friends. I wanted my old friends*", where nothing indicates the author, except "Elena" in the fourth sentence, which can be considered as the author "I".

### 2.6 Comparing CDS and Referents

This part of our system has been partially developed for Entrance Exams evaluation, due to the specificities of this evaluation, specially the triple structure text/questions/answers. Once each text analyzed, each question is analysed, then the four possible answers are analyzed. The questions have generally no anaphora or these anaphors refer to words in the question, but the system needs to consider that "the author" (or, sometimes, "the writer") is "I" in the text. Anaphors in questions are very common and the referents are in the answer (rarely) or in the question (more commonly). For example, in the answer "*Because she did not have any pictures of herself*", the pronoun "*she*" refers to "*Margaret*" in the question "*Why didn't Margaret*

*want the author to see her picture while she was alive*?" and "*herself*" refers to "*she*" which refers to "*Margaret*".

When the question is analyzed, besides the CDS structures, the system extracts the type of the question like in our Question Answering system. In Entrance Exams, these types are always non-factual types like cause ("*What made the author decide to have a pen pal in a foreign country?*"), sentiment ("*How did the author feel when he saw Margaret's photograph?*"), aim ("*Why did the author ask Margaret for her picture*?"), signification ("*By they make up for lost time, the author means that the rats*"), event ("*What happened regarding the house in the end?*") and so on. Frequently, parts of the question need to be integrated into the answers. In the last sentence, for example, the nominal group "*the rats*" needs to be added at the beginning of the answers. In this case, first answer "*come to enjoy their life without friends to play with*" will become "*the rats come to enjoy their life without friends to play with*".

Once the CDS and the type are extracted of the question, referents and temporal and spatial contexts (if they can be extracted from the question) are used to define the part of the text where elements of the answer are the most probable. For example, in the third text where is the precedent question about "*the rats",* this noun appears only in the second half of the text, so the target of the answers is the second half, not the first one, i.e. CDS of the second half will weight more than CDS from the first half and CDS with rats (the noun or an anaphora referring to this noun) will weight more.

In a first time, the system eliminates answers where there is no correspondence between CDS, referents and type of question/answer. There are very few cases, only 7 on 224 answers. More generally, it seems that the method consisting to reduce the choices between answers by elimination of inadequate answers is extremely difficult to implement. Because, probably, answers are made to test the comprehension of the texts and, frequently, the answer which seems to be the best choice (i.e. which integrates the bigger number of words from the text) is not the good one... and, reciprocally, the answer which seems the farest is frequently the good one !

For the answers, two tasks are very important: adding eventually part of the question (described above) and resolution of anaphora. Hopefully the resolution of anaphora is easiest on question and answers than in the text. The number of possible referents is reduced and, testing on the evaluation run, we found that the system only made two errors : in "*make it easier for older workers to acquire new skills*" with the question "*Changes in technology can*", "*it*" is considered to refer to "*technology*" when, here, we have a cataphora and "*it*" refers to "*acquire new skills*". And in the answer "*Tom was to shine a coloured torch onto Jenny's face to make it look horrible*", "*it*" is given as referring to "*torch*" when it refers to "*Jenny's face*".

Equivalences between the subject "I" and a proper noun is not so frequent in the evaluation test as it is in the training corpus. But this equivalence is not so evident for the text 23 (next to last) where this equivalence needs to be deducted from: *"I was only seven years old at the time, but I still remember that day. "Elena, we're going to Japan.""* And this equivalence is very important because "Elena" is the subject of four questions out of five!

To compare CDS of answers and CDS of text, we compare each CDS of text to each CDS of each answer, taking into account a coefficient of proximity of the target

and the number of common elements. Subject and verb have bigger weight than object, direct or indirect, which have bigger weight than temporal and spatial context. If the system finds two elements in common, the total is multiplied by 4, if three elements are in common, the total is multiplied by 16, etc. The system also increases the total when there is a correspondence with the type of the question. If only one element or no element is common to the CDS, the system takes into account the categories of our ontology, increasing the total if there is a correspondence. The total is slightly increasing if there are common referents. The total is cumulative with all the CDS of the text and finally divided by the number of CDS in the answer (often one, no more than three in the evaluation corpus).

At the end, we have, for each answer, a coefficient which ranges from 0 to 32792 (in the evaluation test, because there is no upper limit). The answer with the biggest coefficient is considered as the correct answer.

## 3 Results

Our system answered correctly to 25 questions out of 56 (c@1 = 0.45). The $\chi^2$ is 11.52 (i.e. a probability of 0,09% that these results were obtained randomly). Knowing that, randomly, a system will obtain an average 25% of good answers, in this case 14 good answers. Thus, we outperform random only from 11 good answers, which is not a good result because it means that all our syntactic and semantic methods perform only an improvement of 11 answers out of 42 (total of 56 questions decreased of 14 due to random). Even if this result is the second best, underperforming our results for French language, we cannot consider that our main hypothesis is verified. It seems clear for us that, without pragmatic knowledge and natural language inference, it's impossible to obtain more than 0.6, like we obtained for French.

However the score difference between French and English runs suggests that it's possible to improve English results if we use similar resources and modules. Currently, our company is improving its English parser. Nevertheless, in the version used for this evaluation, a bug caused the phrasal verbs not to be taken into account (we discover that after the end of the evaluation!). And our resolution of anaphora is also presently less successful in English than in French and so is the detection of the type of the question. So, in all the areas, we need to improve the English modules to obtain similar results for the two languages. And that is what we are doing until the end of 2014.

With the run results files, we tested different hypothesis. In a first hypothesis (see Figure 2, Results with different filters for answers). We keep only answers where the probability of the best answer is superior or equal to 1000. In this case, we have 9 good answers on 16 questions. Even if the percentage of success is 56%, in fact the c@1 is equal to 0,276, which is lower than the result on 56 questions. If we keep only the questions where the probability of the best answer is superior or equal to 500, we obtain 16 good answers on 28. In this case, results are better: the percentage of success is 57% and the c@1 is equal to 0,429, very close to our result of 0,446 on the total of questions. Finally we keep only the answers where the probability of the best

answer is almost twice the probability of the second best answer. In this case, we obtain 9 good answers on 19, which is the worst result with 47% of successful answers but a c@1is equal to 0.267, a little bit more than random!

| | Results | % successful | c@1 |
|---|---|---|---|
| evaluation run | 25/56 | 45 % | 0.45 |
| probability >= 1000 | 9/16 | 56 % | 0.28 |
| probability >= 500 | 16/28 | 57 % | 0.43 |
| best >= 2nd best | 9/19 | 47 % | 0.27 |

Figure 2. Results with different filters for answers.

So, in all the cases, our c@1 is inferior to 0.5 and our English system will not pass the Entrance Exams for the Japanese University! If we look to the results text by text, on the 12 texts, 7 are superior or equal to 50%.

But there is an area where the computer is clearly superior to the human: speed. The English run is executed in 2.3 seconds, which means a speed of about 3500 words by second. Because we did not try to optimize the code, this speed could be better (the speed of our parser is more than 10000 words by second), specially if we rewrite the comparison between CDS of text and CDS of answers.

## 4 Analysis of results

Last year [1] [2] [10] [15], like this year, there were 5 participants, but only 10 runs (29 runs this year). On these 10 runs, 3 obtain results superior to random and 7 inferior or equal to random. This year, out of 29 runs, 14 obtain results superior to random and 15 inferior or equal to random. If we consider as a good result needs to be independant of chance with a probability higher than 95%, the $\chi^2$ needs to be superior or equal to 3.84. Last year, only one run had a $\chi^2$ superior to 3.84, this year only four runs have a $\chi^2$ superior to 3.84.

These calculations demonstrate the difficulty of the task. The fact that more than half of the runs, this year and last year, obtained results inferior or equal to random, shows that classical methods used in Question Answering don't work on these comprehension reading tests. These tests have been written by humans to evaluate the reading comprehension of humans. So, for example, the answer which seems the best, i.e. which includes the higher number of words from the text, is generally a bad answer.

To demonstrate that with our run, we will take two examples, the first one is very basic, the second one is more complex. As you can imagine, our system finds the good answer in the first case, not in the second case. The easiest question/answer is extracted from text 16:

*What was the man with glasses doing at the barber's when the writer met him?*

*1. He was cutting his hair.*

*2. He was standing in line outside.*

*3. He was talking with other people.*

*4. He was waiting for a haircut.*

Some words in the question like "glasses" or "barber" indicate that the target is at about 10% of the text, with the sentences: *Take the man I met at a barber's in Chicago, for instance. He was last in line waiting for a haircut, and he stared at me through his thick glasses as I walked in and sat next to him*.

Even with a "bag of words" method, the answer 4 can be found as the good one, considering the correspondence "*waiting for a haircut*". A simple resolution of anaphora indicates that the subject "*he*" is "*man meet at a barber's*", so the coefficient of confidence becomes very high. For this question, the coefficient of the answer 4 is 824, which represents more than the triple of the answer 2.

The second example is considerably more complex and our system didn't find the good answer. It is the first question of the text 22:

*Why did Mrs. Tortino agree to the offer from the man in the bowler hat?*

*1. He promised her more sunshine without offering her any money.*

*2. He said they would build a house which looked just like her old one.*

*3. He told her that she would not have to move out of her old house.*

*4. He told her to move to a new building located at the same address.*

The words "man in the bowler hat" indicate a target at about 30% of the text, with the sentences: *Then one day in early spring, a man in a bowler hat came to her door. Somehow he seemed different from the others as he walked all around her shaded house, gazing at the long shadows in the garden and sniffing the foul air. At first Mrs. Tortino thought he would offer her money for her home, like all the rest of the men. But when he began to speak she listened, her eyes opening wide. "Could you really do that?" she asked. The man nodded. "A tall building right where my house stands, but you won't destroy....?" "That's right" he said. "Your house will be under the same sky, on the same street, at the same address. You'll keep everything just as it is. Even Pursifur." "And there will be money for more tomato plants and some flower seeds and cat food for Pursifur?" "Indeed," said the man, smiling. Mrs. Tortino stared at the man in the bowler hat for a long time. Then, at last, she said, "All right!" And they shook hands.*

To answer the question, in fact next sentences are needed but we keep here only sentences which are at the target. As you can read, many facts are implied in the text. To choose the good answer (3 for this question), you need to know that if a house is in the same street and at the same address, then there is no moving... except if you need to go from a house in a building (answer 4). You also need to know that saying "*all right*" and "*shake hands*" is similar to "*agree to the offer*". Our system returned the answer 1, especially because "*there will be money for more tomato plants*" was not considering as contradictory with "*without offering any money*".

## 5 Conclusions

All the software modules and linguistic resources used in this evaluation exist since many years and are the property of the company Synapse Développement. The parts developed for this evaluation are the Machine Reading infrastructure, some improve-

ments of the resolution of anaphora in English and the complete module to compare CDS from text and answers. No external resources or natural language inference engine have been used.

With 25 good answers on 56 questions, the results seem good and this run is the best run for English, the second one in the evaluation after our run in French. The difference of performance for these two languages indicates clearly that we can improve modules for English, probably in all the domains (parsing, word sense disambiguation, resolution of anaphora, searching type of question). But, like for French, the limitations of the method appear clearly: to obtain more than 2/3 of good answers, pragmatic knowledge and inference are essential.

**Acknowledgements**. We acknowledge the support of the CHIST-ERA project "READERS Evaluation And Development of Reading Systems" (2012-2016) funded by ANR in France (ANR-12-CHRI-0004) and realized with the collaboration of Universidad del Pais Vasco, Universidad Nacional de Educación a Distancia and University of Edinburgh. This work benefited from numerous exchanges and discussions with these partners led within the framework of the project.

## 6 References

1. Arthur, P., Neubig, G., Sakti, S., Toda, T., Nakamura, S., NAIST at the CLEF 2013 QA4MRE Pilot Task. *CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5,* ISSN 2038-4963, Valencia - Spain, 23 - 26 September, 2013 (2013)
2. Banerjee, S., Bhaskar, P., Pakray, P., Bandyopadhyay, S., Gelbukh, A., Multiple Choice Question (MCQ) Answering System for Entrance Examination, Question Answering System for QA4MRE@CLEF 2013. *CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5, ISSN 2038-4963,* Valencia - Spain, 23 - 26 September, 2013 (2013)
3. Buck, G., Testing Listening Comprehension in Japanese University Entrance Examinations, *JALT Journal, Vol. 10, Nos. 1 & 2*, 1988 (1988)
4. Iftene, A., Moruz, A., Ignat, E.: Using Anaphora resolution in a Question Answering system for Machine Reading Evaluation. *Notebook Paper for the CLEF 2013 LABs Workshop* - QA4MRE, 23-26 September, Valencia, Spain (2013)
5. Indiana University, French Grammar and Reading Comprehension Test. *http://www.indiana.edu/~best/bweb3/french-grammar-and-reading-comprehension-test/*
6. Kobayashi, M., An Investigation of method effects on reading comprehension test performance, *The Interface Between Interlanguage, Pragmatics and Assessment: Proceedings of the 3rd Annual JALT Pan-SIG Conference*. May 22-23, 2004. Tokyo, Japan: Tokyo Keizai University (2004)
7. Laurent, D., Séguéla, P., Nègre, S., Cross Lingual Question Answering using QRISTAL for CLEF 2005 *Working Notes, CLEF Cross-Language Evalua-*

*tion Forum, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, 20-22 september 2006, Alicante, Spain (2006)
8. Laurent, D., Séguéla, P., Nègre, S., Cross Lingual Question Answering using QRISTAL for CLEF 2006, Evaluation of Multilingual and Multi-Modal Information Retrieval Lecture Notes in Computer Science, *Springer, Volume 4730*, 2007, pp 339-350 (2007)
9. Laurent, D., Séguéla, P., Nègre, S., Cross Lingual Question Answering using QRISTAL for CLEF 2007, *Working Notes, CLEF Cross-Language Evaluation Forum, 7th Workshop of the Cross-Language Evaluation Forum*, CLEF 2008, Budapest, Hungary (2008)
10. Li, X., Ran, T., Nguyen, N.L.T., Miyao, Y., Aizawa, A., Question Answering System for Entrance Exams in QA4MRE. *CLEF 2013 Evaluation Labs and Workshop Online Working Notes, ISBN 978-88-904810-5-5,* ISSN 2038-4963, Valencia - Spain, 23 - 26 September, 2013 (2013)
11. MacCartney, B., Natural Language Inference, PhD Thesis, Stanford University, June 2009 (2009)
12. Mulvey, B., A Myth of Influence: Japanese University Entrance Exams and Their Effect on Junior and Senior High School Reading Pedagogy, JALT Journal, Vol. 21, 1, 1999 (1999)
13. National Institute of Informatics, Todai Robot Project, NII Today, n°46, July 2013 (2013)
14. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE at CLEF 2012: Question Answering for Machine Reading Evaluation. CLEF 2012 *Evaluation Labs and Workshop Working Notes Papers*, 17-20 September, 2012, Rome, Italy (2012)
15. Peñas, A., Miyao, Y., Hovy, E., Forner, P., Kando, N. : Overview of QA4MRE at CLEF 2013 Entrance Exams Task. *CLEF 2013 Evaluation Labs and Workshop. Online Working Notes, ISBN 978-88-904810-5-5*. ISSN 2038-4963 (2013)
16. Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Evaluating Machine Reading Systems through Comprehension Tests. *LREC 2012 Proceedings of the Eight International Conference on Language Resources and Evaluation*, 21-27 May, 2012, Istambul (2012)
17. Peñas, A., Rodrigo, Á. : A Simple Measure to Assess Non-response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1415–1424*, Portland, Oregon, June 19-24, 2011. Association for Computational Linguistics (2011)
18. Quintard, L., Galibert, O., Adda, G., Grau, B., Laurent, D., Moriceau, V., Rosset, S., Tannier, X., Vilant, A. , Question Answering on Web Data : The QA Evaluation in Quero, *Proceedings of the Seventh Conference on Language Resources and Evaluation*, 17-23 May, 2010, Valletta, Malta (2010)
19. Riloff, E., Thelen, M., A Rule-based Question Answering System for Reading Comprehension Tests. Proceedings of ANL P/NAACL 2000. Workshop on Reading Comprehension Tests as Evaluation for computer-Based Language Understanding Systems, PP. 13-19 (2000)