Evaluating and Improving the Distractor-Generating Heuristics

Itziar Aldabe, Montse Maritxalar (1), Edurne Martinez
Department of Computer Languages and Systems
University of the Basque Country
P.O. box 649, E-20080 Donostia
(1) montse.maritxalar@ehu.es

Abstract

ArikIturri is a system developed for the automatic generation of didactic resources. More specifically, it is focused on the automatic generation of questions based on NLP tools.

In this paper, we present an evaluation of the error correction and multiple-choice question types generated by ArikIturri. This evaluation has been carried out with the collaboration of four expert editors.

In the mentioned question types, heuristics have been used in order to generate the distractors automatically. Although the heuristics have been manually defined, a first attempt for their automatic generation is explained.

Keywords

question generation, evaluation, heuristics

1 Introduction

In the last years, several research on automatic generation of questions for language learning has been carried out. Among others, reading comprehension, vocabulary, cloze questions and grammar tests are automatically generated based on Natural Language Processing (NLP) techniques and resources. These tests deal with different topics related to specific linguistic phenomena or to more general subjects such as the comprehension of a text. We call topic to the subject matter of a question.

The evaluation of the systems can also be made from different perspectives. For example, [8] evaluates the system in terms of time (to produce distractors) and quality of the items. The questions previously approved by a linguistic lecturer are evaluated by students. [6] also evaluates the system giving the cloze tests to 60 students. [7] makes use of authoring and assessment subsystems in order to evaluate the generated questions. In [3] the vocabulary questions are compared to human-generated questions, while in [4] the grammar questions are evaluated by seven professor and students. Finally, [9] examines the quality of the questions with the help of a native speaker of English.

In this paper, we compare the opinions of different human editors about the question automatically generated by ArikIturri. Firstly, the questions created by the generator are evaluated by a single editor, and then, the evaluation is extended to some more editors in order to compare their points of view. Moreover, the improvements of distractor-generating heuristics is also the aim of our work. A distractor is a choice which does not match correctly in the context of the question and a heuristic is the rule or the knowledge the system uses to generate the distractors. The heuristics have been defined manually whereas the distractors are automatically generated words. In this paper, we also explain a first attempt to generate the heuristics automatically.

Section 2 presents the question generator. In section 3, the editors' agreement is analysed. Section 4 deals with different ways of producing distractors. Finally, some conclusions and future work are outlined.

2 The Question Generator

ArikIturri [1] is a system developed for the automatic generation of didactic resources based on NLP tools. The system generates different types of questions using pedagogical corpora. Although we have developed it for the Basque language, the architecture of the system is language independent.

The input corpus consists of a databank which is composed of morphologically and syntactically analysed sentences where phrase chunks are automatically identified. Question instances of a question model consist the output of the system. Both input and output are represented in XML.

Figure 1 represents the automatic process for question generation.

The *sentence retriever* module in ArikIturri selects candidate sentences from the source corpus. In a first step, the candidate sentences for the questions are automatically extracted from the databank, depending on the topic of the question. Then, it analyses the occurrences rate of the possible candidates in order to make random selection of the sentences.

Once the sentences are selected, the *answer focuses identificator* tags the chunked phrases (answer focuses) where the topic to be treated appears. Then, the *item generator* creates the questions depending on the specified exercise type. That is why this module contains the *distractor generator* submodule. Distractors are automatically generated words; they are not extracted from any databank. Due to the rich inflection system

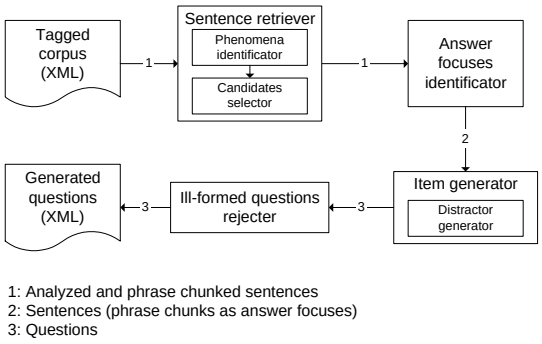

Fig. 1: *Architecture of ArikIturri*

of Basque, it is impossible to store every possible word form in a dictionary, even in a compressed way. Therefore, we use a general purpose morphological generator to create the distractors. By contrast, the heuristics used by the generator are not automatically generated, they are based on experts' knowledge. In section 4 we will explain some experiments carried out for the automatic generation of the heuristics.

As the question generation process is automatic, it is probable that some of the questions are ill-formed. This is why we have included the *ill-formed questions rejecter* in the architecture of our system.

One of the aims in [1] was to prove the viability of ArikIturri when constructing questions. Even though the system was able to produce four different types of questions, we limited our evaluation to multiple-choice (500 questions) and error correction (1,200 questions) question types. The system used 17 heuristics for the generation of the distractors for these questions.

100 sentences were selected at random for each heuristic from a high language level corpus of 10,079 sentences[1]. Therefore, the evaluation was carried out with 1,700 sentences. Moreover, the experiment was limited to two types of linguistic phenomena. On the one hand, the following declension cases were the topic of the questions: dative (DAT), sociative (SOZ), inessive (INE), ergative (ERG) and absolutive (ABS). On the other hand, the present and past indicative verb tenses were chosen for the experiment.

Table 1 shows the different heuristics that the system used to generate the distractors. The heuristics were defined taking into account common mistakes that students make when learning Basque such as the use of wrong declension cases or finiteness. As regards the verb tenses, the heuristics change the different persons of the verb that belong to the different auxiliary paradigms[2].

The generator (by means of the *ill-formed questions rejecter* module) automatically rejected 58 multiple-choice and 292 error correction instances out of all the generated questions[3]. In some cases the system creates the same two distractors or rejects a badly-formed distractor, etc. This way, a sample of 1,350 question instances was obtained to be evaluated manually.

Declension cases	Change of the finiteness	Sociative
		Inessive
	Replacement of declension cases	SOZ => ABS
		SOZ => DAT
		SOZ => ERG
		INE => ABS
		INE => DAT
		INE => ERG
		ABS => SOZ
		ABS => INE
		DAT => SOZ
		DAT => INE
		ERG => SOZ
		ERG => INE
Verb	Change of the person of the verb	DA paradigm
		DU paradigm
		ZAIO paradigm

Table 1: *Heuristics*

3 Editors' agreement

One way of evaluating the questions generated by ArikIturri is to give them to different editors. In this section, we present two experiments carried out for the manual evaluation of ArikIturri.

It is necessary to underline the fact that the heuristics used for the generation of the distractors (see table 1) were based on the knowledge of an expert who took part in the design of the system but not in the evaluation. It is also important to remark that the heuristics were defined all in a row by a single person.

3.1 Experimental settings

Four experts in both computational linguistics and language teaching took part in the two experiments as editors: three of them work as computational linguists in a NLP research group and the last one in HABE (Institute for the Teaching of Basque and Basque Language Literacy to Adults), which is an institute of the Basque government developed for L2 and L1 Basque Language Teaching. The computational linguists do not only have a linguistic profile but also a language teaching background. In the case of the expert language teacher from HABE, he has teaching profile as well as experience in creating didactic resources.

For the two experiments, the editors used a web-based post-editing environment which helped them to set the questions. In the first experiment, the manual evaluation was carried out by a computational linguist while in the second experiment three new editors took part in it. The aim of these experiments was to compare the opinions of different editors.

The sample of 1,350 question instances mentioned in section 2 was evaluated in the first experiment. One expert evaluated 442 multiple-choice questions and 908 of the error correction type. In the case of multiple-choice, the topic of 153 questions was the declension case and the other 289 questions had the verb tense

[1] The input corpus was classified into three different language levels chosen by expert teachers.
[2] The paradigms are explained in section 4.1.
[3] We did not evaluate them manually because they were badly-formed questions.

as topic. However, in the case of the error correction questions, they all were connected with declension cases.

In the second experiment, we took 100% of the generated multiple-choice questions related to the present indicative verb tense and 25% of the error correction questions related to declension cases. The sample we obtained contained a total of 431 questions nearly the same amount for each linguistic phenomenon.

Table 2 summarises the information of the experimental settings.

	1st experiment	**2nd experiment**
Number of Editors	1	3
Multiple-choice Declension cases	153	0
Multiple-choice Verb tenses	289	195
Error correction Declension cases	908	236
Total amount of questions	1350	431

Table 2: *Experimental settings*

3.2 First experiment

In the evaluation of the first experiment, we asked the editor to modify or reject questions only if they were badly-formed. Considering that all the questions discarded or modified by the editor were not well generated, the results showed that the rate of the accepted questions was 82.71% in the case of error correction questions and 83.26% in the case of multiple-choice questions.

Let us assume that the probability to generate a proper distractor and consequently a theoretically acceptable question in error correction is $P(dist) = p$ where $0 \leq p \leq 1$. However, the probability to generate an acceptable question decreases in multiple-choice questions (3 distractors in this case). If we assume that the distractors are independent with each other, the probability to create an acceptable question is $P(dist_1, dist_2, dist_3) = P(dist_1) \cdot P(dist_2) \cdot P(dist_3) = p \times p \times p$ where $0 \leq 3p \leq p \leq 1$. The results obtained in the experiment confirm that this probability has influence on the acceptance rate, when dealing with the same topic. For instance, in the case of declension cases the acceptance rate is 82.71% for error correction question type and 64.70% for multiple-choice.

If we split the multiple-choice questions taking into account the number of distractors we find that there is a significant difference in terms of acceptance. Regarding the verb tenses, the acceptance rate is 92.73%, while in the case of declension cases, it is 64.70%. As the system generated two distractors when dealing with verb tenses and three when dealing with declension cases, the probability of creating a correct question for verbs is higher than in the case of declension cases. However, the acceptance rate given by the human editors (92.73%) in the case of verb tenses (2 distractors) is higher than the rate (82.71%) for the declension cases (1 distractor). Table 3 displays all the acceptance rates in this first experiment.

	Topic	Number of distractors	Acceptance rate
Error correction	Declension cases	1	82.71%
Multiple-choice	Declension cases AND Verb tense	2-3	83.26%
	Declension cases	3	64.70%
	Verb tenses	2	92.73%

Table 3: *Accepted questions*

We conclude that the number of distractors alter the acceptance rate of the generated questions by ArikIturri. We foresaw that the topic could also have influence in the results.

3.3 Second experiment

In order to corroborate the results from the first experiment as well as our hypothesis about the topic, three new editors took part in the second experiment. Two of them were asked to evaluate the accepted questions in the first experiment. The first one had to evaluate the questions related to verb tenses, the second one those related to declension cases, and the third editor examined the questions which were rejected.

	Verb	Declension cases
Accepted in the 1st	94.97%	96.94%
Rejected in the 1st	75.00%	25.00%

Table 4: *Evaluation of the 431 questions*

Table 4 shows the results we obtained in the second experiment in comparison with the ones obtained in the first one. For example, the editor of the second experiment accepted 94.97% of the questions related to the verb tenses which were previously accepted in the first experiment and 96.94% of the declension cases. For instance, both editors agreed on the fact that the error correction question "**Industriak milaka profesionalek galdu du entzumena enpresetako zarataren erruz*[4]" is an acceptable question because the correct answer is "Indrustian[5]" instead of "Industriak". 75% of the rejected questions related to verb tenses in the first experiment were also not accepted in the second one, while in the case of declension cases the percentage is 25%.

A more detailed information is given in table 5, where the number of questions in which different editors agree and disagree are displayed.

[4] *The industry lots of professionals have lost their hearing due to the factory noise

[5] In the industry lots of professionals have lost their hearing due to the factory noise

Declension cases		
	Accepted in 2	Rejected in 2
Accepted in 1	190	6
Rejected in 1	30	10
Verb tenses		
	Accepted in 2	Rejected in 2
Accepted in 1	170	9
Rejected in 1	4	12

Table 5: *Comparison of the results of the two experiments*

Both tables show good results, in fact, the second experiment also verifies the high percentage of well-formed questions. The favourable opinion of the four editors is also an important aspect since the questions were automatically generated.

However, although it is supposed that all the editors followed the same instructions for the evaluation of the automatically generated questions, we must also consider other aspects, such as chance or some personal factors, which might have influence on the results obtained in the evaluation. Those factors could be i) editors' own experience when generating questions manually; ii) the final users of the questions they are thinking about; iii) when and how the evaluation was carried out; iv) the number of questions to evaluate, etc. Cohen's kappa index (κ) [5] takes into account this variable.

If we apply the kappa concept to our results, we obtain the kappa indexes displayed in table 6.

	Kappa
Declension cases Error correction	0.28
Verb tenses Multiple-choice	0.61
Total	0.4

Table 6: *Editors' agreement (kappa)*

If we take into account that there are more distractors in a multiple-choice question than in an error correction question, the probability that two editors will agree is higher in the case of error correction. In the case of multiple-choice questions, they must agree on all the different distractors. Therefore, as the number of questions evaluated for each question type was almost the same, we should expect better kappa indexes in the case of declension cases since they belong to the error correction type of question. As a consequence, we conclude that the topic of the questions have influence on the results. Indeed, our hypothesis when planning the second experiment was that it was easier to generate questions to learn verb tenses than to learn declension cases.

4 Producing distractors

We have already mentioned the fact that ArikIturri makes use of different heuristics to create the distractors of the questions. The information used to define the heuristics was manually created but it could be also automatically generated. The rules represent some of the unsuitable combinations from the linguist point of view.

As our aim is also to generate the heuristics automatically, in the next sections, we explain the first attempts we have made in this research line.

4.1 Automatic extraction of patterns to define heuristics

The previous experiments have been performed having human experience on the basis of the implementation of the heuristic rules. The new approach considers an automatic process not only for the generation of the distractors but also for the generation of the heuristics.

Before focusing on the automatic extraction of patterns to define the heuristics, we consider necessary to clarify two aspects:

- *Verb and ellipsis:* The auxiliary verbs in Basque refer to the syntagmatic component where the ergative, absolutive and dative cases occur. Even if those phrases do not appear in the sentence, the auxiliary gives us that information and we can know which phrases have been elided.

  In general, a verb can have from one to four different auxiliary paradigms. These paradigms correspond to the following four auxiliary types:

  - DA: the absolutive is the subject of the clause.
  - DU: the ergative is the subject and the absolutive is the direct object of the clause.
  - DIO: the ergative is the subject, the absolutive is the direct object and the dative is the indirect object of the clause.
  - ZAIO: the absolutive is the subject and the dative is the indirect object of the clause.

- *Working unit:* In this article, the term *clause* refers to a group of phrases containing a conjugated verb. A sentence that contains only one clause is called a simple sentence; if we have two or more sentences (juxtaposition, coordination or subordination), we speak of complex sentences. Therefore, we consider two different working units: the simple sentence level and the complex sentence level.

The basis of the automatic extraction of patterns to define heuristics comes from [2], where a finite-state syntactic grammar was developed in order to join the verb instances and their corresponding syntactic dependents (arguments and adjuncts) from journalistic corpora. The grammar scores 87% of precision and 66% of recall. The system obtained 688 different patterns for 640 verbs. For each verb more than one of the different 688 patterns can occur.

The patterns which represent the knowledge extracted from automatically analysed corpora were obtained at simple sentence level. Moreover, they automatically retrieved the elided cases in order to reflect

them in the patterns. Each of the patterns offers the following information: i) the syntactic dependents; ii) the auxiliary type and iii) the number of instances. For instance, one of the 143 extracted patterns related to the verb "askatu" (to release) is:

48 askatu: DU: ABS + ERG + INE

Based on the journalistic corpora, the system developed in [2] matched 48 times the DU: ABS + ERG + INE pattern for the verb "askatu" (to release), that is, it reflects the number of times that the absolutive, the ergative and the inessive occur with the auxiliary DU.

We consider these patterns could be the basis for the automatic generation of the heuristics. When the morphological generator creates a distractor, the patterns automatically extracted and the created distractor pattern could be compared. If a matching was detected, the distractor could not be considered a candidate distractor and the question would be automatically rejected.

The experiments previously carried out offered us the chance to compare the heuristics manually generated with the automatic patterns. Moreover, as the questions were already manually evaluated, we could study the measure of success of the patterns.

Once the clauses of the questions are extracted, it is important to specify which phrases are going to be taken into consideration when matching them to the patterns. For example, considering that the system has generated the following error correction question when setting absolutive as the topic, we will study different criteria for comparing the automatically extracted patterns with the phrases of the question generated by ArikIturri:

"*Hainbat ariketaren bidez *gure gorputzaren* ***blokeoarekin*** askatu dugu"[6]

The phrase containing the correct answer ("gure gorputzaren blokeoa[7]") is absolutive and it has been transformed into the sociative (SOZ) case in order to generate the distractor ("gure gorputzaren blokeoarekin"). The phrase "Hainbat ariketaren bidez" refers to the instrumental case (INS) and the auxiliary for the verb "askatu" is DU.

The auxiliary DU for the verb "askatu" tells us there is a subject (ERG) as well as a direct object (ABS).

The criteria to compare the clause of the question with the patterns can be summarised as follows:

- *Criterion 1*: to compare the patterns with the declension cases/phrases that appear in the clause explicitly. In the previous example, in the case of the distractor, DU: INS + SOZ would be compared with the patterns from [2]. As there is no matching, ArikIturri would create a distractor.

If we want to take into account the phrases containing some given declension cases that occur in a clause plus those which are elided, we can follow two different options:

- *Criterion 2*: to contrast the cases which have been elided, if they are not part of the topic. In the example, the system would compare DU: INS + SOZ + ERG with the automatic patterns. That is to say, we would take into consideration the ergative case because it is elided and it is not the topic. On the contrary, we would not consider the absolutive case because it is the topic of the question. In this case the system would generate a distractor, since it does not match any of the 143 patterns extracted for the verb "askatu".
- *Criterion 3*: to include all the elided cases. In that example, we would compare the paradigm DU: INS + SOZ + ERG + ABS with the patterns. As this distractor pattern exists in [2], the system would not generate a distractor.

In the next two sections, we explain the two approaches followed in the automatic generation of the heuristics. The first one was developed to generate complex sentence questions while the second one was carried out to create simple sentence questions.

4.2 Heuristics based on patterns to generate complex sentence questions

The first attempt was to compare the evaluated questions in the first experiment with the patterns automatically extracted from journalistic corpora. For that, some steps were followed:

1. To obtain a sample of the error correction questions related to the sociative, inessive, ergative, dative or absolutive cases. This was the same sample as the one used for the second experiment (25% of error correction questions).
2. To extract the simple sentence of each question where the topic appeared. This task was handmade. When the topic was part of the subordinate clause, the subordinate clause was manually transformed into a main clause.
3. To compare the questions (at clause level) with the patterns in order to observe the acceptance rate if, in the basis, we had an automatic generation of heuristics based on the automatic patterns.

As editors agreed in a high rate in the previous experiments, we first made a study of the heuristics used in the generation of well-formed questions. We compared the questions accepted in the first experiment with the patterns. That is to say, the information of the well-formed questions was divided to compare both the correct answer and the distractors.

If we applied the *criterion 1*, we might expect low results since it only takes into account the explicit phrases of the question, while the patterns automatically assign the explicit declension cases of the verb as well as those that are elliptic. Nevertheless, low results are obtained in the case of the correct answers, but not in the case of the distractors.

Table 7 shows the results related to the three different criteria when comparing both data of the questions accepted in the first experiment:

For instance, if we had, in the basis, the automatically extracted knowledge (patterns) when using *criterion 1*, 66.27 out of the 100 correct answers accepted

[6] *(we) have released ***with*** *the stiffening of our body* by means of some exercises.

[7] the stiffening of our body.

	Criterion 1	Criterion 2	Criterion 3
Correct answer in the 1st experiment	66.27%	93.59%	93.59%
Distractor in the 1st experiment	69.94%	66.46%	37.89%

Table 7: *Accepted questions*

by the editor in the first experiment would be also considered correct answers. Besides, 69.94% refers to the clauses of the questions that were accepted by the editor as distractors. Almost 70% of the distractors of the questions would also be considered distractors if the patterns were used for the automatic generation of the heuristics.

Regarding the results obtained in the case of the *criterion 2*, we could conclude that they are more realistic and better. As the correct answers were extracted from the source sentence of the corpus, they are presumably correct.

The number of the questions created by ArikIturri and rejected by the editors is not troubling, since it is a low percentage[8]. However, the patterns can also be compared with them in order to observe if the distractors rejected by the human editor would not be created having the patterns as basis of the heuristics.

Table 8 represents the percentages related to the rejected distractors in the first experiment together with the three different criteria.

	Criterion 1	Criterion 2	Criterion 3
NOT distractor in the 1st experiment	15.38%	19.23%	48.00%

Table 8: *Rejected questions*

As it happens in the case of the accepted questions, in this case the way to compare the questions with the patterns is threefold. This time the given percentages have a different meaning. Since the questions were rejected by the editor, it is supposed that the rejected distractors were not proper ones[9]. If we used the automatically extracted patterns to create heuristics and compare them only with the explicit phrases of the clause, 15.38% of the distractors would be considered improper distractors in the case of *criterion 1.* In this case, better results are obtained from the third comparison.

We foresaw some aspects that could affect the results: the error rate of the patterns, the corpus and the working unit. The error rate of the patterns may alter the results since their precision is 87%. Therefore, 13% of the times the patterns obtained in [2] are incorrect. As regards the corpus, it could also be an aspect to be considered because different corpora have been used in both works. In the case of ArikIturri, the corpus is focused on language learning while in the automatic extraction of patterns the corpus is composed of newspaper texts. Finally, the experiments commented in section 3 were carried out using complex sentences while the extracted patterns refer to simple ones. Therefore, the working unit could have influenced on the results.

[8] 17.29% in the first experiment for error correction question types and 6.78% in the second one

[9] The error correction question type has just one distractor.

4.3 Heuristics based on patterns to generate simple sentence questions

The fact that the working unit could affect in the results made us carry out a new experiment in order to obtain heuristics based on patterns to generate simple sentence questions. This time, we presented the editors new questions to be evaluated. Those questions were simple sentences manually extracted from the complex sentences used in the previous experiments. As we have said, we used a sample of 25% of the error correction questions which were related to the sociative, inessive, ergative, dative and absolutive cases.

The editors evaluated the questions and accepted 75.21% of them. If we compare it with the acceptance rate in the first (82.71%) and second (93.22%) experiments, the acceptance rate decreases. These results correspond to the generated error correction question types related to the declension cases. The only difference lies in the evaluated sentences: the ones from the first and second experiment were complex sentences, whereas in the last one they were simple sentences.

Once we obtained a set of questions manually evaluated, we compared them with the automatically generated patterns. This time, we used the three criteria previously mentioned again. Table 9 shows the patterns accuracy taking into account the accepted questions (75.21%).

	Criterion 1	Criterion 2	Criterion 3
Correct in the 1st experiment	100%	100%	100%
Distractor in the 1st experiment	79.19%	78.05%	77.64%

Table 9: *Accepted questions at simple sentence level*

The results from the experiment are better than those shown in table 7. In all cases, the patterns would consider all the correct answers. Regarding the distractors, the best results are obtained from the first comparison, although there is not a significant difference between the three evaluations. Moreover, the results are closer to the error rate of the automatically extracted patterns than in table 7.

In the case of the rejected questions, the same equivalence was carried out. Table 10 displays the obtained results.

	Criterion 1	Criterion 2	Criterion 3
NOT distractor in the 1st experiment	38.46%	38.46%	40.00%

Table 10: *Rejected questions at simple sentence level*

Although the results are better than the ones obtained in table 8, they are still quite poor. In case of the *criterion 1* and *criterion 2* the results are twice better, but still poor.

The comparison of the results obtained in the evaluations show us that a clause considered a question at complex sentence level is not always a question at simple sentence level, and vice versa. Moreover, the editors took into account the elided sentence elements when evaluating the questions at simple sentence level. Finally, we also conclude that the best results are obtained from *criterion 2*.

5 Conclusions and future work

In this paper, we have presented an automatic question generator, which was evaluated in different ways. First, one editor evaluated multiple-choice and error correction question types created by ArikIturri. The error correction questions were generated to deal with declension cases, and the multiple-choice questions cope with declension cases as well as with verb tenses. Apart from obtaining a high acceptance rate in the automatically generated questions, we can conclude that automatically generate error correction questions are more reliable than the multiple-choice ones.

After that, the evaluation was extended to three new editors in order to compare the different opinions they might have. In all cases, the number of the accepted questions was high, and so it was the agreement among the editors. Moreover, based on the obtained results, we can also conclude that the topic has influence in the automatic generation of questions.

In the generation process of the questions that were manually evaluated by different editors, the distractors were automatically generated, but the heuristics were manually defined. Nevertheless, one of the purposes of this work was to try to automatize this process. Section 4 deals with it, comparing the patterns automatically extracted from journalistic corpora with the patterns of complex sentence questions and simple sentence questions. The automatic generation of heuristics to create distractors is more reliable if we choose simple sentences as questions instead of complex ones. However, the acceptance rate of the same questions is higher in the case of complex sentences: 82.71% in the first experiment, 93.22% in the second experiment with complex sentences and 75.21% in the last experiment with simple sentences.

As future work, we plan to continue working on these heuristics. We must not only to try to automatize the process but also study their combination. We plan to analyse different possibilities such splitting the heuristics taking into account the learning level, combining different heuristics to deal with the same topic, etc. Finally, as regards the evaluation carried out, we consider necessary to extend it to learners in order to obtain more realistic results.

Acknowledgements. This research is supported by the University of the Basque Country (GIU05/52) and the Ministry of Industry of the Basque Government (ANHITZ project, IE06-185).

References

[1] I. Aldabe, M. L. de Lacalle, M. Maritxalar, E. Martinez, and L. Uria. Arikiturri: An automatic question generator based on corpora and nlp techniques. *Proceedings of the Eigth International Conference on Intelligent Tutoring Systems (ITS'06)*, pages 584–594, 2006.

[2] I. Aldezabal, M. Aranzabe, A. Atutxa, K. Gojenola, M. Oronoz, and K. Sarasola. Application of finite-state transducers to the acquisition of verb subcategorization information. *Natural Language Engineering. Cambridge University Press.*, 9(1):39–48, 2003.

[3] J. Brown, G. Frishkoff, and M. Eskenazi. Automatic Question Generation for Vocabulary Assessment. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 819–826, 2005.

[4] C. Chen, H. Liou, and J. Chang. FAST - An Automatic Generation System for Grammar Tests. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 1–4, 2006.

[5] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, pages 37–46, 1960.

[6] D. Coniam. A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests. *CALICO Journal*, 16(2–4):15–33, 1997.

[7] C. Liu, C. Wang, Z. Gao, and S. Huang. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, pages 1–8, 2005.

[8] R. Mitkov, L. Ha, and N. Karamis. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering: Special Issue on using NLP for Educational Applications*, 12(2):177–194, 2006.

[9] E. Sumita, F. Sugaya, and S. Yamamoto. Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, pages 61–68, 2005.