Multiple Choice Question Corpus Analysis for Distractor Characterization

V-M. Pho [*] [†], **T. André** [*] [•], **A-L. Ligozat** [*] [⋈], **B. Grau** [*] [⋈], **G. Illouz** [*] [†], **T. François** [•]
[*]LIMSI-CNRS, Orsay, France, firstname.lastname@limsi.fr
[†] Université Paris-Sud, Orsay, France
[⋈] ENSIIE, Evry, France
[•]CENTAL, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgique

Abstract

In this paper, we present a study of MCQ aiming to define criteria in order to automatically select distractors. We are aiming to show that distractor editing follows rules like syntactic and semantic homogeneity according to associated answer, and the possibility to automatically identify this homogeneity. Manual analysis shows that homogeneity rule is respected to edit distractors and automatic analysis shows the possibility to reproduce these criteria. These ones can be used in future works to automatically select distractors, with the combination of other criteria.

Keywords: MCQ, corpus analysis, syntactic and semantic homogeneity

1. Introduction

Technology enhanced learning environments have developed over the past years. In order to be widely used by students and teachers, they must provide means for self-evaluation, and assistance to teachers for generating exercises. We focus on Natural Language Processing (NLP) methods for a particular aspect of such environments, which is Multiple Choice Question (MCQ) generation. Creating a MCQ requires to give clues to generate the question, the correct answer, and some incorrect answers, called distractors.

The aim of this study is to identify the characteristics of distractors in MCQs from a comprehensive corpus-based analysis of MCQs, in order to be able to generate them automatically.

2. Problem definition

A MCQ is composed of two parts (see Figure 1): the *stem*, which will be called *question* in the rest of the paper, since it often takes the form of a question, and the *alternatives*, which include both the correct answer, and one or several *distractors* (incorrect answers).

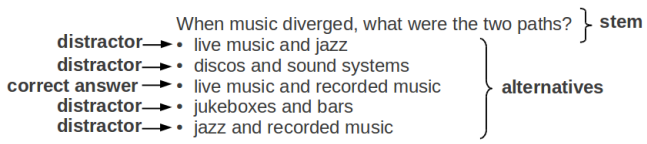

Figure 1: MCQ example (from QA4MRE 2011)

A difficulty when creating a MCQ is to choose distractors (Rodriguez, 2005): the quality of the MCQ relies on it. Many guidelines have been proposed for distractor generation, by teachers and education experts (for example see (Burton et al., 1991) or (Leclercq, 1986)). These guidelines provide general recommendations based on high level criteria, without defining them precisely. In our work, instead of relying on a set of general rules to evaluate quality distractors, we established a set of characteristics from a corpus analysis to obtain computable criteria.

Following the guidelines, we identified two main notions in the corpus: homogeneity (lexical, semantic, syntactic) between the alternatives for a same question; and specificity, i.e. what differentiates a distractor from the correct answer. Regarding the homogeneity, we studied two main characteristics of distractors with respect to the correct answer: syntactic homogeneity and semantic homogeneity. Syntactic homogeneity requires that distractors share (at least partially) a common syntactic structure with the answer (except for special distractors such as "other", "all of the above" etc.). For example in Figure 1, all alternatives are composed of a coordination of noun phrases. Semantic homogeneity states that alternatives share a common semantic type (expected by the question). For the example in Figure 1, all alternatives are musical styles or types.

These two notions meet the kind of criteria used by works including distractor generation, such as (Mitkov et al., 2006; Karamanis et al., 2006; Lee and Seneff, 2007; Zweig and Burges, 2012). The first two base their selection on semantic resources (resp. WordNet and UMLS) to select concepts with the same hypernym for distractors; the last two are specialized in fill-in-the-blank questions: (Lee and Seneff, 2007) use corpus frequencies and (Zweig and Burges, 2012) select distractors from an N-gram model.

In our work, we define characteristics of distractors which are domain and language independent, and apply to all types of answers. In order to study the validity of these characteristics and the possibility to automate their recognition, we tested them on different types of MCQs, first manually, and then with automatic annotations.

3. Corpus and annotations

3.1. Corpus

We collected a corpus of MCQs varying according to the task that has to be solved: (i) Tests for evaluating human knowledge of a domain, (ii) Tests for evaluating human level of language understanding and (iii) Tests for evaluating machine abilities to understand a text (Machine reading tests). Our purpose was to study which characteristics of the descriptors are shared among the different types of MCQs and which of them are different.

Machine reading MCQs (qa4mre in Table 1) are provided by the evaluation campaign QA4MRE 2012 (Peñas et al., 2013) for the main task. Each MCQ corresponds to questions about a given document. The other English MCQs were extracted from different websites, either related to domain knowledge evaluation (mcq1) or to language evaluation (mcq2). In mcq2, questions also correspond to a given document. The topics of these MCQs are miscellaneous. Table 1 summarizes the characteristics of the corpus.

corpus	lang	#q	#a	purpose	topic
qa4mre	en	100	500	Mach. read.	Alzheimer, music/society, climate change, AIDS
mcq2	en	68	298	lang. eval.	English understanding
mcq1	en	20	80	knowl. eval.	Theology/law, history/geography

Table 1: Characteristics of the corpus: MCQ name, language, number of questions, number of answers, purpose and topic

3.2. Manual annotation

3.2.1. Annotation process

The corpus was built by three people. One of them annotated the corpus, and the last two checked the annotations. For the annotation process, we used Brat (Brat Rapid Annotation Tool) (Stenetorp et al., 2012).

3.2.2. Annotation categories

In order to validate the syntactic and semantic homogeneity hypothesis, we manually annotated the distractors in the corpus. Syntactic annotation aims at comparing syntactic structures between a distractor and the correct answer, independently of the meaning, whereas semantic annotation aims at analyzing conformity between the distractor and the correct answer.

Syntactic annotations Correct answers can be expressed using different syntactic structures, often related to the question form: named entity, phrase (noun phrase or verbal phrase), clause or sentence. Thus, we defined syntactic criteria that make it possible to compare these different syntactic structures. Distractors are classified into four categories:

- **identical syntax:** Represents distractors that have the same chunk sequences as that of the correct answer. Chunks are defined here as the lowest-level phrases of the constituency parse tree covering the entire sentence [1]. For example, the distractor "The number of tortoises began to decrease."

  "**NP**(The number) **PP**(of tortoises) **VP**(began) **VP**(to decrease)."

  and the correct answer "The number of tortoises began to grow."

  "**NP**(The number) **PP**(of tortoises) **VP**(began) **VP**(to grow)."

  have an identical syntax: their chunk sequence is "*NP PP VP VP*";

- **partially identical syntax:** Represents distractors that share the same chunk sequence as their associated answers, but for one variation (chunk insertion, deletion or substitution). For example, the distractor "it resists diseases"

  "**NP**(it) **VP**(resists) **NP**(diseases)"

  and the correct answer "it is not profitable"

  "**NP**(it) **VP**(is not) **ADJP**(profitable)"

  have one variation: the last ADJP of the correct answer is substituted by a NP in the distractor;

- **globally identical syntax:** Represents distractors which have more than one chunk variation according to the correct answer, but share the same global structure as the correct answer, i.e. same kinds of clauses or same kinds of high-level phrases. For example, the distractor "because the amount of CO2 saved by using renewable energies is not considered" and the correct answer "because they only consider current emissions but not previous ones" differ by more than one variation, but they both are subordinate causal clauses: the syntax is considered globally identical;

- **different syntax:** Represents distractors that do not share the same global syntax with the correct answer. For example, the distractor "military operations and migrant labor" and the correct answer "leveraging financial funds and financing HIV/AIDS programs for Africa" have a different syntax: the distractor is a coordination of noun phrases whereas the correct answer is a coordination of verbal phrases.

Semantic annotations At the semantic level, descriptors have a different meaning from the answer meaning. However, a certain semantic conformity can be found for example in relation to the expected type of answer (EAT) deduced from the question. This type can be a *specific type*, given explicitly in the question (for example "Which president had the most children?" expects as answer a person who is a "president"); or a *named entity type* (for example "Who invented the telephone?" expects a named entity of the *Person* type for an answer); or a semantic role (the question "Why do patients in Africa have an almost total lack of access to ARV drugs?" expects a *reason* as an answer). The homogeneity can also be studied in relation with the correct answer type (AT), i.e. the theoretical named entity type of the answer. Thus, we defined two kinds of manual semantic

[1] We made an exception for prepositional phrases containing only a preposition and a nominal chunk which are considered as a single prepositional chunk (for example "in the laws" will be considered as a prepositional chunk).

annotations of alternatives on the one hand, and distractors on the other hand.
The first type of annotation determines whether the alternatives correspond to the EAT (deduced by the annotator, and not the result of a question analysis module):

- **conform type:** The descriptor type is conform to the most precise EAT. The precise EAT is a specific type if it exists. If not, it is a named entity type. If the answer type is not given, but only its semantic relation (causes, definitions...), the descriptor is considered to be of a conform type if it constitutes a possible argument for this role;
- **non conform type:** Includes alternatives whose type is different from the EAT;
- **unknown conformity**: Represents alternatives for which it is impossible to evaluate the conformity according to the EAT, i.e. alternatives for which the annotator cannot identify the type or for which it is impossible to identify the EAT of their associated questions (for example, "When using a file...").

The second type of annotation determines whether the distractors have the same named entity type as the answer (w.r.t. QALC system named entity taxonomy (Ferret et al., 2000)). Our analysis is restricted to named entities because there is no reference hierarchy. As in the previous annotation, named entity types are speculated by the annotator:

- **identical named entity type:** Represents distractors that share their named entity type with the correct answer;
- **different named entity type:** Represents distractors that do not share their named entity type with the correct answer;
- **not a named entity:** Represents distractors that are not named entities.

We manually annotated each distractor of the corpus by attributing them one category for each kind of annotation.

3.3. Automatic annotations

To complete manual annotations, we also analyzed the corpus with NLP tools. The objective was to test whether automatic annotations also enabled to verify the syntactic and semantic homogeneity, and thus could be later used for distractor generation.

3.3.1. Syntactic comparison

For the syntactic homogeneity, we compared the constituency-based parse trees of the answers and of the distractors, at different levels: part-of-speech tagging, chunk level, parse tree top-level and full tree.
To compare the automatic annotations with the manual ones, we computed the Levenshtein distance between the sequence of chunks of the distractor and that of the correct answer.

- If the distance is 0, we consider that the distractor and the correct answer have an identical syntax;
- If it is 1 (cost of an operation), we consider that they have a partially identical syntax;
- If the distance is greater than 1 and that the answer contains complete clauses, we compare the top-level nodes of the parse trees. If the distractor and the correct answer share the same sequence of nodes at this level, we consider that they have a globally identical syntax;
- If not, they have a different syntax.

To check if the lengths of the distractors influence syntactic homogeneity, we also compared the sequence of the chunks, and computed a tree edit distance on the entire tree (Zhang and Shasha, 1989).
The parsing of answers and distractors was performed by the Stanford Parser (Klein and Manning, 2003).

3.3.2. Semantic comparison

For the semantic homogeneity, we limited our analysis to named entity types. To perform the semantic analysis, we used the Stanford Named Entity Recognizer (Finkel et al., 2005). We considered the following named entity types annotated by this tool: Time, Location, Organization, Person, Money, Percent, Date, Duration, Ordinal, Set and Miscellaneous. Although the types are different from those used for the manual annotation, the purpose is similar namely to compare the named entity type of the answers and distractors. We also compared the named entity type of the distractors and the EAT derived from the question analysis (Ligozat, 2013).

4. Corpus analysis

We report here the main results of our corpus study.

4.1. Results of the manual annotation

4.1.1. Syntactic annotations

We annotated 479 distractors with syntactic information from the 650 distractors of the corpus (eliminating redundant examples). Table 2 shows the distribution of these distractors in the syntactic categories.

	Number	Percentage
Identical syntax	189	39.5 %
Partially identical syntax	91	19 %
Globally identical syntax	141	29.4 %
Different syntax	58	12.1 %
Total	479	100 %

Table 2: Results of the syntactic manual annotation

According to these results, we observe that about 40 % of annotated distractors share a common syntax for their associated answers. These distractors are mostly named entities but some sentences and clauses belong to this category. The remaining distractors are mainly the result of a verb or subject substitution from their associated answers structures.

Half of distractors categorized as "partially identical syntax" are lists or distractors which their associated answers are lists, like the following example:

> **Distractor:** Union and State List
> **Correct answer:** Concurrent List, Union List, Residuary Subject List

Moreover, a part of distractors belonging to this category can indeed present small syntactic variations with the correct answer, but remain quite similar.
Almost all distractors categorized as "globally identical syntax" are clauses. The reason is their syntactic structures do not follow a syntactic homogeneity as strict as distractors like chunks and named entities. The latter category, "different syntax", does not contain a lot of distractors (12 %) and is partly composed of distractors such as "none of the above", which do not contain real possible answers.

4.1.2. Semantic annotations

As regards semantic annotations, we annotated 609 of the 838 answers of the corpus (73 %). Table 3 shows the distribution of these alternatives according to their type conformity to the question EAT.

	Number	Percentage
Conform type	460	75.5 %
Non conform type	26	4.3 %
Unknown conformity	123	20.2 %
Total	609	100 %

Table 3: Results of conformity of the alternative AT to the question EAT

About three-quarters of the alternatives share the same type as the question EAT. This observation shows that conformity of the question EAT is a criterion to select distractors. Nevertheless, we could not identify the conformity of 20 % of the alternatives: the named entity type can not be used to characterize these distractors.
The annotation of correspondence of named entity types between distractors and answers was realized on 484 distractors of the corpus (74 %). Table 4 shows the distribution of these distractors.

	Number	Percentage
Identical named entity type	102	21.1 %
Different named entity type	17	3.5 %
Not a named entity	365	75.4 %
Total	484	100 %

Table 4: Results of the manual annotation on named entity correspondence between answers and distractors

About three-quarters of the distractors are not of a QALC named entity type. Almost all other distractors have the same named entity type as the answer. Table 5 shows the relation between the alternatives types in the case of distractors which do not share the same named entity type as the answer. Over the 17 concerned distractors, 12 of them are related to the expected type (*country* instead of *city* for example). We did not find any distractor whose named entity type was not related to the one of its associated answer.

	Number	Percentage
Hyperonymy	4	23.5 %
Hyponymy	8	47.1 %
Other (different NE)	0	0 %
Other (not a NE)	5	29.4 %
Total	17	100 %

Table 5: Relations between distractors and answers which have different named entity types

We also conducted analyses according to the kinds of MCQs. Tables 6, 7 and 8 show the results of the manual annotations according to the different parts of the corpus (qa4mre, mcq2 and mcq1).

	qa4mre	mcq2	mcq1
Identical syntax	30.4 %	44.7 %	51.7 %
Partially identical synt.	22 %	18 %	16.7 %
Globally identical synt.	38.2 %	23.4 %	23.3 %
Different syntax	9.4 %	14.9 %	8.3 %

Table 6: Results of the syntactic manual annotation according to the different parts of the corpus

	qa4mre	mcq2	mcq1
Conform type	96.8 %	56 %	75.9 %
Non conform type	2.4 %	5.3 %	6 %
Unknown conformity	0.8 %	38.7 %	18.1 %

Table 7: Results of conformity of the alternative AT to the question EAT according to the different parts of the corpus

	qa4mre	mcq2	mcq1
Identical NE type	18 %	19.6 %	33.9 %
Different NE type	6 %	1.8 %	0 %
Not a NE	76 %	78.7 %	66.1 %

Table 8: Results of the manual annotation on named entity correspondence between answers and distractors according to the different parts of the corpus

qa4mre has more questions that explicit the EAT than MCQs dedicated to language evaluation, but the latter have more distractors which share the same named entity type as their associated answers. qa4mre distractors have more often a globally identical syntax than those of the other MCQs, whose syntactic conformity are mostly in the first two categories.

4.2. Results of the automatic annotation and comparison to the manual one

In this sub-section, we present the results of the automatic annotation. To compare them to the manual one, we automatically annotated the same distractors as the manual annotation. In order to evaluate the automatic annotation, we present first the repartition of distractors according to categories and we computed *recall*, *precision* and *f-score* for each category evaluated. Formulae of these metrics are the following:

$$recall(cat) = \frac{\text{\# distractors correctly classified}}{\text{\# distractors automatically classified in } cat}$$

$$precision(cat) = \frac{\text{\# distractors correctly classified}}{\text{\# distractors manually classified in } cat}$$

$$f\text{-}score(cat) = 2.\frac{precision(cat).recall(cat)}{precision(cat) + recall(cat)}$$

We automatically evaluated the syntactic annotation and the named entity correspondence. Apart from these comparisons, we checked if syntactic homogeneity is correlated with lengths of distractors, since the manual annotation tended to show that longer distractors (and their answers) are syntactically less homogeneous.

4.2.1. Syntactic annotations

Table 9 shows the distribution of distractors in the different syntactic categories based on distances in terms of chunk sequences.

	Automatic	Manual
Identical syntax	32.9 %	39.5 %
Partially identical syntax	32.4 %	19 %
Globally identical syntax	25.7 %	29.4 %
Different syntax	9 %	12.1 %

Table 9: Results of the syntactic automatic annotation and comparison to the manual one

We observe that percentages of the automatic annotation are quite similar to the manual one, except that a lower proportion of distractors was labeled as "identical syntax" and a higher proportion of them was labeled as "partially identical syntax". However, the distractors classified in one or the other class are different, as Table 10 shows.

	Precision	Recall	F-score
Identical synt.	0.71	0.83	0.77
Partially identical synt.	0.50	0.30	0.38
Globally identical synt.	0.58	0.67	0.62
Different synt.	0.28	0.36	0.31

Table 10: Precision, recall and f-scores of the syntactic automatic annotation

We observe that distractors with a syntax identical to their answers have largely been recognized by the automatic annotation, and more than half of the distractors with a globally identical syntax to their answers have been recognized. However, distractors with a partially identical or different syntax from their answers are not very well recognized. A part of these distractors are manually classified as "globally identical syntax" and automatically annotated as "partially identical syntax" or vice versa. A less important part of distractors manually annotated as "partially identical syntax" are automatically annotated as "identical syntax" because their identified chunk sequence (from parse trees) are similar, as the following example illustrates:

Distractor:
The sailors use the tortoises for food.
NP(The sailors) **VP**(use) **NP**(the tortoises) **PP**(for food).
Answer:
The scientists raise the tortoises in special pens.
NP(The scientists) **VP**(raise) **NP**(the tortoises) **PP**(in special pens).

More than half of distractors with a different syntax have been recognized as "partially identical syntax" due to the fact that the automatic annotation only takes into account chunks and top-level nodes, and not the whole syntactic structure.

Distractor:
the installation of hydro power plants
NP(the installation) **PP**(of hydro power plants)
Answer:
spontaneous fires
NP(spontaneous fires)

In this example, we observe that syntactic structures are different, but the top-level node of both sentences is "NP". Figure 2 shows the Levenshtein distance between chunks of distractors and answers according to lengths of these distractors.

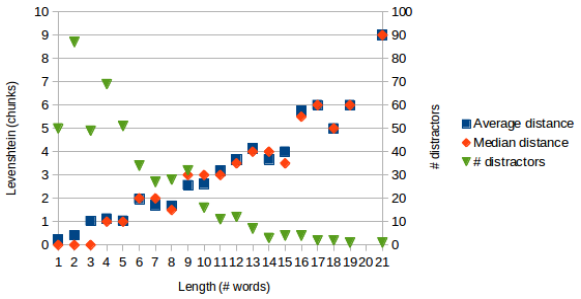

Figure 2: Levenshtein distance between chunks of distractors and their associated answers according to lengths of distractors, and number of distractors per length

This figure shows that a large part of distractors are small (less than 10 words) and that very small distractors (between 1 and 3 words) generally have no chunk variation, according to their associated answers. In order to analyse the other distractors, we took into account the average number of words per chunk (about 2 words). We observe that distractors composed of more than 3 words have generally a

partial variation according to their associated answers (according to the average number of words per chunks, between 1/3 and 2/3 of the words). According to Table 9, these distractors have a partially or globally identical syntax as their associated answers.
Figure 3 shows the tree edit distance between parse trees of distractors and their associated answers according to lengths of these distractors. In order to compare syntactic structures, we removed leaves (words) of these trees. Compared to the previous analysis which allows to observe variations in the low-level of syntactic structures, the analysis of tree edit distance allows to observe variations on the overall parse trees.

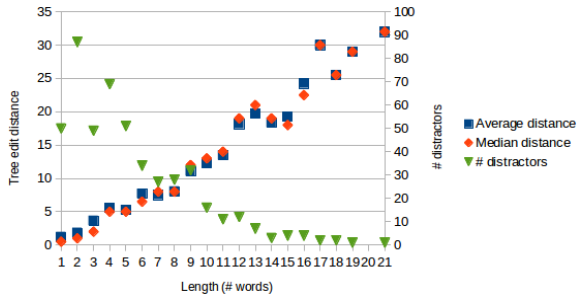

Figure 3: Tree edit distance between parse trees of distractors and their associated answers according to lengths of distractors, and number of distractors per length

We observe that a part of the syntactic structure of the distractors is shared with this of their associated answers. Moreover, like the analysis of Levenshtein distances between distractors and answers, the tree edit distance increases strongly for big distractors (more than 10 words). This shows that syntactic homogeneity is respected to create distractors, especially short distractors like named entities and chunks.

4.2.2. Semantic annotations

As regards semantic annotations, we compared automatic annotations to manual ones in terms of conformity of the named entity type of the alternatives according to the EAT of their associated questions, and in terms of similarity of named entity types of the distractors and their associated answers.
In order to compare the annotations related to the EAT of questions, we do not compare alternatives which are manually classified as "unknown conformity" because they do not provide a base to evaluate the automatic annotation. Table 11 show the repartition of the alternatives in the two remaining categories.

	Automatic	Manual
Conform type	62.4 %	94.7 %
Non conform type	37.5 %	5.3 %

Table 11: Results of the automatic annotation on conformity of alternatives compared to the EAT of the question, and comparison to the manual annotation

We observe a significant difference between the two annotations: manual annotation found that almost all alternatives have a conform type according to the EAT of their associated questions, whereas the automatic annotation found just over one half of the alternatives in this category. Table 12 shows performance of this automatic annotation.

	Precision	Recall	F-score
Conform type	0.97	0.64	0.77
Non confom type	0.10	0.69	0.17

Table 12: Precision, recall and f-scores of the automatic annotation on EAT relations

We observe that alternatives classified as "conform type" are relatively well classified. However, it is not the case for the other category. We identified three main reasons causing these bad results: the question analyzer fails to recognize the type of non-interrogative questions (for example, the question "Yoshiko is in New York City because..." is recognized as a question expecting a *location*, whereas it expects a reason). The other reasons are due to mislabelings of the named entity recognizer or alternatives which are not named entities. The two following items shows these two phenomena:

> **Question:** For how long has Rebecca Lolosoli been working with MADRE? (*type: duration*)
> **Distractor:** since the late 1990s (*correct type: duration, type labeled by the named entity recognizer: date*)

> **Question:** Where does Yoshiko's adventure begin? (*type: location*)
> **Distractor:** at the TeenSay offices (*correct type: location but not annotated by the named entity recognizer because it is not a named entity*)

Concerning the correspondence between the types of the distractors and the answer, Table 13 shows the repartition of distractors in terms of named entity type correspondence.

	Automatic	Manual
Identical named entity type	13.1 %	21.1 %
Different named entity type	4.5 %	3.5 %
Not a named entity	82.4 %	75.4 %

Table 13: Results of the automatic annotation on named entity type of distractors compared to this of the answer, and comparison to the manual annotation

In comparison to the manual annotation, the automatic one found more distractors which are not named entities and less with an named entity type similar to the answer. Table 14 shows performance of automatic annotation of named entity homogeneity between distractors and their associated answers.

	Precision	Recall	F-score
Identical NE type	0.94	0.61	0.74
Different NE type	0.32	0.44	0.37
Not a NE	0.92	1.00	0.96

Table 14: Precision, recall and f-scores of the automatic annotation on named entity relations

We observe that almost all distractors without named entity type are well recognized. Moreover, a large part of distractors with the same named entity type as their associated answers are detected and more than half of distractors manually annotated as "identical named entity type" are well recognized. Other ones are often recognized as "not a named entity type" because their named entity type is not a category of Stanford Named Entity Recognizer or are mislabeled, like the distractor "*The Methodist Church*" which is not recognized as an *Organization*. Distractors which share a different named entity type with their associated answers are not well classified: one third of them are recognized as "identical named entity type" and another third are recognized as "not a named entity".

4.3. Discussion

As part of our analysis, we observed that it is possible to manually classify distractors in categories representing syntactic and semantic homogeneity between them and their associated answers. We note that redaction rules formulated by (Burton et al., 1991) at the syntactic and semantic levels are respected. At the syntactic level, about 40 % of distractors have the same syntax as their associated answers, and if we take into account distractors which present an almost identical syntax as their answer, the proportion increases to about 90 %. At the semantic level, we note that about 75 % of distractors and answers present the expected type of their questions and about 85 % of distractors share the same named entity type as their associated answers, in the case of named entities answers.

The automatic annotations gave distributions similar to the manual ones. The results of the syntactic annotation shows the possibility to automatically represent syntactic homogeneity between the distractors and their associated answers. The Levenshtein distance between chunk sequences and the tree edit distance show that syntactic variations between the distractors and their associated answers are partial and are low in the case of small distractors. Moreover, the semantic annotations presented in this paper showed that it is possible to automatically represent a partial semantic homogeneity, even if the verification of the distractor type can be improved taking into account specific types provided by knowledge bases.

These results show that it is possible to apply the methods that we presented in this paper to characterize distractors respecting (Burton et al., 1991)'s methodology. Even if these methods must be extended to other criteria, it is possible to extract fragments from texts according to semantic and syntactic characteristics of a base answer, and transform these fragments into distractors. To complete semantic homogeneity recognition, we can take into account other criteria coming for example from semantic relatedness recognition methods.

5. Conclusion and future works

In this paper, we presented a MCQ corpus-based study aiming at identifying the characteristics of distractors. We first approached the issue manually in order to check our hypotheses and then automatically to check the possibility to automatically generate distractors. Manual annotations showed that distractors are largely homogeneous, from the syntactic and semantic points of view. Automatic annotations return correct results, especially the annotation related to named entities which recognize 70 % of relations between named entities of distractors and their associated answers. The syntactic annotation recognize half of the relations between parse trees of distractors and their associated answers, especially distractors which share the same syntax than their associated answers. The annotation related to the EAT of questions properly recognize alternatives whose named entity types are conform to the EAT of their associated questions.

In a first time, semantic specificity will be further explored. Then, these characteristics of distractors will be used to create a distractor validation system, and be integrated into a MCQ generation system. Moreover, we will adapt these methods in other languages like French.

6. Acknowledgments

This work has been partly financed by Digiteo under the Aneth project.

7. References

Burton, S. J., Sudweeks, R. R., Merrill, P. F., and Wood, B. (1991). *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., Masson, N., and Lecuyer, P. (2000). QALC–The Question-Answering System of LIMSI-CNRS. In *TREC 9 Notebook*.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Karamanis, N., Ha, L. A., and Mitkov, R. (2006). Generating multiple-choice test items from medical text: A pilot study. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 111–113. Association for Computational Linguistics.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Leclercq, D. (1986). *La conception des questions à choix multiple*. Collection EDUCATION 2000. Editions LABOR.

Lee, J. and Seneff, S. (2007). Automatic generation of cloze items for prepositions. In *INTERSPEECH*, pages 2173–2176.

Ligozat, A.-L. (2013). Question classification transfer. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 429–433, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mitkov, R., Ha, L. A., and Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194.

Peñas, A., Hovy, E., Forner, P., Rodrigo, Á., Sutcliffe, R., and Morante, R. (2013). QA4MRE 2011-2013: Overview of Question Answering for Machine Reading Evaluation. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 303–320. Springer.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2):3–13.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

Zweig, G. and Burges, C. J. (2012). A challenge set for advancing language modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 29–36. Association for Computational Linguistics.