

Generating educational assessment items from Linked Open Data: the case of DBpedia

Muriel Foulonneau

Tudor Research Centre,
29, av. John F. Kennedy
L-1855 Luxembourg, Luxembourg
muriel.foulonneau@tudor.lu

Abstract. This work uses Linked Open Data for the generation of educational assessment items. We describe the streamline to create variables and populate simple choice item models using the IMS-QTI standard. The generated items were then imported in an assessment platform. Five item models were tested. They allowed identifying the main challenges to improve the usability of Linked Data sources to support the generation of formative assessment items, in particular data quality issues and the identification of relevant sub-graphs for the generation of item variables.

Keywords: Linked Data, open data, DBpedia, eLearning, e-assessment, formative assessment, assessment item generation, data quality, IMS-QTI

1 Introduction

Assessment takes a very important role in education. Tests are created to evaluate what students have learned in the class, to assess their level at the beginning of a cycle, to enter a prestigious university, or even to obtain a degree. More and more assessment is also praised for its contribution to the learning process through formative assessment (i.e., assessment to learn, not to measure) and/or self-assessment whereby the concept of a third party controlling the acquisition of knowledge is totally taken out of the assessment process. The role of assessment in the learning process has considerably widened. The New York Times even recently published an article entitled “To Really Learn, Quit Studying and Take a Test” [1], reporting on a study by Karpicke et al. [2] which suggests that tests are actually the most efficient knowledge acquisition method.

The development of e-assessment has been hampered by a number of obstacles, in particular the time and effort necessary to create assessment items (i.e., test questions) [3]. Therefore, automatic or semi-automatic item generation has gained attention over the last years. Item generation consists in using an item model and creating automatically or semi-automatically multiple items from that model.

The Semantic Web can provide relevant resources for the generation of assessment items because it includes models of factual knowledge and structured datasets for the

generation of item model variables. Moreover, it can provide links to relevant learning resources, through the interlinking between different data sources.

Using a heterogeneous factbase for supporting the learning process however raises issues related for instance to the potential disparities of data quality. We implemented a streamline to generate simple choice items from DBpedia. Our work aims at identifying the potential difficulties and the feasibility of using Linked Open Data to generate items for low stake assessment, in this case formative assessment.

We present existing approaches to the creation of item variables, the construction of the assessment item creation streamline, and the experimentation of the process to generate five sets of items.

2 Existing work

Item generation consists in creating multiple instances of items based on an item model. The item model defines variables, i.e., the parts which change for each item generated. There are different approaches to the generation of variables, depending on the type of items under consideration.

In order to fill item variables for mathematics or science, the creation of computational models is the easiest solution. Other systems use natural language processing (NLP) to generate for instance vocabulary questions and cloze questions (fill in blanks) in language learning formative assessment exercises ([4], [5], [6]). Karamanis et al. [7] also extract questions from medical texts.

The generation of variables from structured datasets has been experimented in particular in the domain of language learning. Lin et al. [8] and Brown et al. [9] for instance generated vocabulary questions from the WordNet dataset, which is now available as RDF data on the Semantic Web. Indeed, the semantic representation of data can help extracting relevant variables. Sung et al. [10] use natural language processing to extract semantic networks from a text and then generate English comprehension items.

Linnebank et al. [11] use a domain model as the basis for the generation of entire items. This approach requires experts to elicit knowledge in specifically dedicated models. However, the knowledge happens to already exist in many data sources (e.g., scientific datasets), contributed by many different experts who would probably never gather in long modeling exercises. Those modeling exercises would have to be repeated over time, as the knowledge of different disciplines evolves. Moreover, in many domains, the classic curricula, for which models could potentially be developed and maintained by authorities, are not suitable. This is the case of professional knowledge for instance.

Given the potential complexity of the models for generating item variables, Liu [12] defines reusable components of the generation of items (including the heuristics behind the creation of math variables for instance). Our work complements this approach by including the connection to semantic datasets as sources of variables. Existing approaches to item generation usually focus on language learning [13] or mathematics and physics where variable can be created from formulae [14]. We aim to define approaches applicable in a wider range of domains (e.g., history) by reusing existing interlinked datasets.

3 Generating item variables from a SPARQL endpoint

An item model includes a stem, options, and potentially auxiliary information [15]. Only the stem (i.e., the question) is mandatory. Response options are provided in the case of a multiple choice item. Auxiliary information can be a multimedia resource for instance. In some cases, other parameters can be adapted, including the feedback provided to candidates after they answer the item.

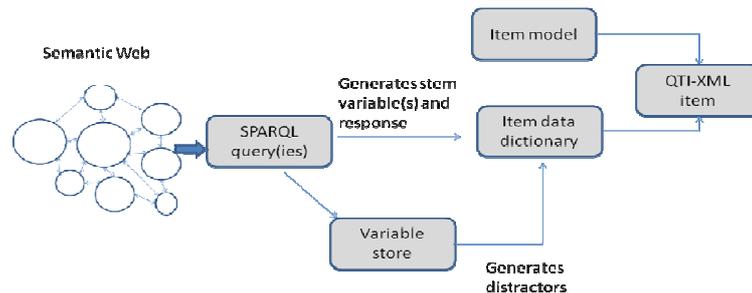


Figure 1 - Semi-automatic item generation from semantic datasets

In order to investigate the use of Linked Data as a source of assessment items, we built a streamline to generate simple choice items from a SPARQL endpoint on the Web. The item generation process is split in different steps detailed in this section. Figure 1 shows the item model represented as an item template, the queries to extract data from the Semantic Web, the generation of a set of potential variables as a variable store, the organization of all the values of variables for each item in data dictionaries, and the creation of items in QTI-XML format from the item template and item data dictionaries. These steps are detailed in this section.

3.1 Creating an IMS QTI-XML template

In order to generate items which are portable to multiple platforms, it is necessary to format them in IMS-QTI (IMS Question & Test Interoperability Specification)¹. IMS-QTI is the main standard used to represent assessment items [16]. It specifies metadata (as a Learning Object Metadata profile), usage data (including psychometric indicators), as well as the structure of items, tests, and tests sections. It allows representing multimedia resources in a test. IMS-QTI has an XML serialization.

¹ <http://www.imsglobal.org/question/>

```

<choiceInteraction responseIdentifier="RESPONSE" shuffle="false" maxChoices="1">
  <prompt>What is the capital of {prompt}?</prompt>
  <simpleChoice identifier="{responseCode1}">{responseOption1}</simpleChoice>
  <simpleChoice identifier="{responseCode2}">{responseOption2}</simpleChoice>
  <simpleChoice identifier="{responseCode3}">{responseOption3}</simpleChoice>
</choiceInteraction>

```

Figure 2 - Extract of the QTI-XML template for a simple choice item

No language exists for assessment item templates. We therefore used the syntax of JSON templates for an XML-QTI file (Figure 2). All variables are represented with the variable name in curly brackets. Unlike RDF and XML template languages, JSON templates can define variables for an unstructured part of text in a structured document. For instance, in Figure 2, the {prompt} variable is only defined in part of the content of the <prompt> XML element. Therefore, the question itself can be stored in the item model, only the relevant part of the question is represented as a variable.

3.2 Collecting structured data from the Semantic Web

In order to generate values for the variables defined in the item template, data sources from the Semantic Web are used. The Semantic Web contains data formatted as RDF. Datasets can be interlinked in order to complement for instance the knowledge about a given resource. They can be accessed through browsing, through data dumps, or through a SPARQL interface made available by the data provider. For this experiment, we used the DBpedia SPARQL query interface (Figure 3). The query results only provide a variable store from which items can be generated. All the response options are then extracted from the variable store (Figure 1).

```

SELECT ?country ?capital
WHERE {
  ?c <http://dbpedia.org/property/commonName> ?country .
  ?c <http://dbpedia.org/property/capital> ?capital
}
LIMIT 30

```

Figure 3 - SPARQL query to generate capitals in Europe

Linked data resources are represented by URIs. However, the display of variables in an assessment item requires finding a suitable label for each concept. In the case presented on Figure 3, the ?c variable represents the resource as identified by a URI. The <http://dbpedia.org/property/commonName> property allows finding a suitable label for the country. Since the range of the <http://dbpedia.org/property/capital> property is a literal, it is not necessary to find a distinct label.

The label is however not located in the same property in all datasets and for all resources. In the example of Figure 3, we used the property `<http://dbpedia.org/property/commonName>` which provides the capital names as literals. However, other properties, such as `<foaf:name>` are used for the same purpose. In any case, the items always need to be generated from a path in a semantic graph rather than from a single triple. This makes Linked Data of particular relevance since the datasets can complete each other.

3.3 Generating item distractors

The SPARQL queries aim to retrieve statements from which the stem variable and the correct answer are extracted. However, a simple or multiple choice item also needs distractors. Distractors are the incorrect answers presented as options in the items. In the case of Figure 3, the query retrieves different capitals, from which the distractors are randomly selected to generate an item. For instance, the capital of Bulgaria is Sofia. Distractors can be Bucarest and Riga.

3.4 Creating a data dictionary from Linked Data

The application then stores all the variables for the generated items in data dictionaries. Each item is therefore represented natively with this data dictionary. We created data dictionaries as Java objects conceived for the storage of QTI data. We also recorded the data as a JSON data dictionary.

In addition to the variables, the data dictionary includes provenance information, such as the creation date and the data source.

3.5 Generating QTI Items

QTI-XML items are then generated from the variables stored in the data dictionary and the item model formalized as a JSON template. We replaced all the variables defined in the model by the content of the data dictionary. If the stem is a picture, this can be included in the QTI-XML structure as an external link.

4 The DBpedia experiment

In order to validate this process, we experimented the generation of assessment items for five single choice item models. We used DBpedia as the main source of variables. The item models illustrate the different difficulties which can be encountered and help assessing the usability of the Linked Data for the generation of item variables.

4.1 The generation of variables for five item models

Q1 - What is the capital of { Azerbaijan }?

The first item model uses the query presented on Figure 3. This query uses the <http://dbpedia.org/property/> namespace, i.e., the Infobox dataset. This dataset however is not built on top of a consistent ontology. It rather transforms the properties used in Wikipedia infoboxes. Therefore, the quality of the data is a potential issue².

Out of 30 value pairs generated, 3 were not generated for a country (Neuenburg am Rhein, Wain, and Offenburg). For those, the capital was represented by the same literal as the country. Two distinct capitals were found for Swaziland (Mbabane, the administrative capital and Lobamba, the royal and legislative capital). The Congo is identified as a country, whereas it has been split into two distinct countries. Its capital Leopoldville was since renamed Kinshasa. The capital of Sri Lanka is a URI, whereas the range of the capital property is usually a de facto literal. Finally the capital of Nicaragua is represented with display technical instructions “Managua right|20px”. Overall, 7 value pairs out of 30 were deemed defective.

Q2 - Which country is represented by this flag ?

```
SELECT ?flag ?country
WHERE {
  ?c <http://xmlns.com/foaf/0.1/depiction> ?flag .
  ?c <http://dbpedia.org/property/commonName> ?country .
  ?c <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  <http://dbpedia.org/class/yago/EuropeanCountries>
}
LIMIT 30
```

Q2 uses the Infobox dataset to identify the label of the different countries. However, the FOAF ontology also helps identifying the flag of the country and the YAGO (Yet Another Great Ontology) [17] ontology ensures that only European countries are selected. This excludes data which do not represent countries.

Nevertheless, it is more difficult to find flags for non European countries, while ensuring that only countries are selected. Indeed, in the YAGO ontology, <http://dbpedia.org/class/yago/EuropeanCountries> is a subclass of <http://dbpedia.org/class/yago/Country108544813>. But most European countries are not retrieved when querying the dataset with <http://dbpedia.org/class/yago/Country108544813>. Indeed, the SPARQL endpoint does not provide access to inferred triples. It is necessary to perform a set of queries to retrieve relevant subclasses and use them for the generation of variables.

Out of 30 items including pictures of flags used as stimuli, 6 URIs did not resolve to a usable picture (HTTP 404 errors or encoding problem).

² <http://wiki.dbpedia.org/Datasets>

Q3 - Who succeeded to { Charles VII the Victorious } as ruler of France ?

```
SELECT DISTINCT ?kingHR ?successorHR
WHERE {
?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/class/yago/KingsOfFrance> .
?x <http://dbpedia.org/property/name> ?kingHR .
?x <http://dbpedia.org/ontology/successor> ?z .
?z <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/class/yago/KingsOfFrance> .
?z <http://dbpedia.org/property/name> ?successorHR
}
LIMIT 30
```

Q3 uses the YAGO ontology to ensure that the resource retrieved is indeed a king of France. Out of 30 results, one was incorrect (*The three Musketeers*). The query generated duplicates because of the multiple labels associated to each king. The same king was named for instance *Louis IX*, *Saint Louis*, *Saint Louis IX*. Whereas de-duplication is a straight forward process in this case, the risk of inconsistent naming patterns among options of the same item is more difficult to tackle. An item was indeed generated with the following 3 options: *Charles VII the Victorious*, *Charles 09 Of France*, *Louis VII*. They all use a different naming pattern, with or without the king's nickname and with a different numbering pattern.

Q4 - What is the capital of { Argentina }? With feedback

```
SELECT ?countryHR ?capitalHR ?pictureCollection
WHERE {
?country <http://dbpedia.org/property/commonName> ?countryHR .
?country <http://dbpedia.org/property/capital> ?capitalHR .
?country <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/class/yago/EuropeanCountries> .
?country <http://dbpedia.org/property/hasPhotoCollection> ?pictureCollection
}
LIMIT 30
```

The above question is a variation of Q1. It adds a picture collection from a distinct dataset in the response feedback. It uses the YAGO ontology to exclude countries outside Europe and resources which are not countries. A feedback section is added. When the candidate answers the item, he then receives a feedback if the platform allows it. In the feedback, additional information or formative resources can be suggested. Q4 uses the linkage of the DBpedia dataset with the Flickr wrapper dataset. However the Flickr wrapper data source was unavailable when we performed the experiment.

Q5 - Which category does { Asthma } belong to?

```
SELECT DISTINCT ?diseaseName ?category
WHERE {
?x <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Disease> .
?x <http://dbpedia.org/property/meshname> ?diseaseName .
?x <http://purl.org/dc/terms/subject> ?y .
?y <http://www.w3.org/2004/02/skos/core#prefLabel> ?category
}
LIMIT 30
```

Q5 aims to retrieve diseases and their categories. It uses SKOS and Dublin Core properties. The Infobox dataset is only used to find labels. Labels from the MESH vocabularies are even available. Nevertheless, the SKOS concepts are not related to a specific SKOS scheme. Categories retrieved range from *Skeletal disorders to childhood*. For instance, the correct answer to the question on *Obesity* is *childhood*.

4.2 The publication of items on the TAO platform

The TAO platform³ is an open source semantic platform for the creation and delivery of assessment tests and items. It has been used in multiple assessment contexts, including large scale assessment in the PIAAC and PISA surveys of the OECD, diagnostic assessment and formative assessment.

We imported QTI items generated for the different item models in the platform, in order to validate the overall Linked Data based item creation streamline. Figure 4 presents an item generated from Q1 (Figure 3) imported in the TAO platform.

Test capital names Venezuela

What is the capital of Venezuela?

- Bangui
- Caracas
- Port Moresby

Figure 4 - Item preview on the TAO platform

³ <http://www.tao.lu>

5 Data analysis

The experimentation of the streamline was therefore tested with SPARQL queries which use various ontologies and which collect various types of variables. It raised two types of issues for which future work should find relevant solutions: the quality of the data and the relevance of particular statements for the creation of an assessment item.

5.1 Data quality challenges

In our experiment, the chance that an item will have a defective prompt or a defective correct answer is equal to the number of defective variables used for the item creation. Q1 uses the most challenging dataset in terms of data quality. 7 out of 30 questions had a defective prompt or a defective correct answer (23,33%).

The chance that an item will have defective distractors is represented by the following formula, where D is the total number of distractors, $d(V)$ is the number of defective variables and V is the total number of variables:

$$D \times \frac{d(V)}{(V - 1)}$$

We used 2 distractors. Among the items generated from Q1, 10 items had a defective distractor (33,33%). Overall, 16 out of 30 items had neither a defective prompt nor a defective correct answer nor a defective distractor (53,33%).

As a comparison, the items generated from unstructured content (text) that are deemed usable without edit were measured between 3,5% and 5% by Mitkov et al. [18] and between 12% and 21% by Karamanis et al. [7]. The difficulty of generating items from structured sources should be lower. Although a manual selection is necessary in any case, the mechanisms we have implemented can be improved.

The ontology

Q1 used properties from the Infobox dataset, which has no proper underlying ontology. Q1 can therefore be improved by using ontologies provided by DBpedia, as demonstrated by Q2 for which no distractor issue was identified. We present Q1 and Q2 to illustrate this improvement but it should be noted that there is not always a straight equivalent to the properties extracted from the Infobox dataset.

Q5 could be improved either if the dataset would be linked to a more structured knowledge organization system (KOS) or through an algorithm which would verify the nature of the literals provided as a result of the SPARQL query.

The labels

The choice of the label for each concept to be represented in an item is a challenge when concepts are represented by multiple labels (Q4). The selection of labels and their consistency can be ensured by defining representation patterns or by using datasets with consistent labeling practices.

Inaccurate statements

Most statements provided for the experiment are not inaccurate in their original context but they sometimes use properties which are not sufficiently precise for the usage envisioned (e.g., administrative capital). In other cases, the context of validity of the statement is missing (e.g., Leopoldville used to be the capital of a country called Congo). The choice of DBpedia as a starting point can increase this risk in comparison to domain specific data sources provided by scientific institutions for instance. Nevertheless, the Semantic Web raises similar quality challenges as the ones encountered in heterogeneous and distributed data sources [19]. Web 2.0 approaches, as well as the automatic reprocessing of data can help improve the usability of the Semantic Web statements. This requires setting up a traceability mechanism between the RDF paths used for the generation of items and the items generated.

Data linkage

Data linkage clearly raises an issue because of the reliability of the mechanism on different data sources. Q3 provided 6 problematic URIs out of 30 (i.e., 20%). Q4 generated items for which no URI from the linked data set was resolvable since the whole Flickr wrapper data source was unavailable. This clearly makes the generated items unusable. The creation of infrastructure components such as the SPARQL Endpoint status for CKAN⁴ registered data sets⁵ can help provide solutions to this quality issue over the longer run.

Missing inferences

Finally, the SPARQL endpoint does not provide access to inferred triples. Our streamline does not tackle transitive closures on the data consumer side (e.g., through repeated queries), as illustrated with Q3. Further consideration should be given to the provision of data including inferred statements. Alternatively, full datasets could be imported. Inferences could then be performed in order to support the item generation process.

Different strategies can therefore be implemented to cope with data quality issues we encountered. Data publishers can improve the usability of the data, for instance with the implementation of an upper ontology in DBpedia. However, other data quality issues require data consumers to improve their data collection strategy, for instance to collect as much information as possible on the context of validity of the data, whenever it is available.

5.2 Data selection

The experiment also showed that the Linked Data statements should be selected. The suitability of an assessment item for a test delivered to a candidate or a group of candidates is measured in particular through such information as the item difficulty.

⁴ <http://www.ckan.net>

⁵ <http://labs.mondeca.com/sparqlEndpointsStatus/index.html>

The difficulty can be assessed through a thorough calibration process in which the item is given to beta candidates for extracting psychometric indicators. In low stake assessment, however, the evaluation of the difficulty is often manual (candidate or teacher evaluation) or implicit (the performance of previous candidates who took the same item). In the item generation models we have used, each item has a different construct (i.e., it assesses a different knowledge). In this case, the psychometric variables are more difficult to predict [20]. A particular model is necessary to assess the difficulty of items generated from Semantic Web sources. For instance, it is likely that for a European audience, the capital of the Cook Islands will raise a higher rate of failure than the capital of Belgium. There is no information in the datasets, which can support the idea of a higher or lower difficulty. Moreover, the difficulty of the item also depends on the distractors, which in this experiment were generated on a random basis from a set of equivalent instances. As the generation of items from structured Web data sources will become more elaborated, it will therefore be necessary to design a model for predicting the difficulty of generated items.

6 Conclusion and future work

The present experimentation shows the process for generating assessment items and/or assessment variables from Linked Data. The performance of the system in comparison with other approaches shows its potential as a strategy for assessment item generation. It is expected that data linkage can provide relevant content for instance to propose formative resources to candidates who failed an item or to illustrate a concept with a picture published as part of a distinct dataset.

The experimentation shows the quality issues related to the generation of items based on such a resource as DBpedia. It should be noted that the measurements were made with a question which raises particular quality issues. It can be easily improved as shown with other questions. Nevertheless the Linked Data Cloud also contains datasets published by scientific institutions, which may therefore raise less data accuracy concerns. In addition, the usage model we are proposing is centered on low stake assessment, for which we believe that the time saved makes it worthwhile having to clean some of the data, while the overall process remains valuable.

Nevertheless, additional work is necessary both on the data and on the assessment items. The items created demonstrate the complexity of generating item variables for simple assessment items. We aim to investigate the creation of more complex items and the relevance of formative resources which can be included in the item as feedback. Moreover, the Semantic Web can provide knowledge models from which items could be generated. Our work is focused on semi-automatic item generation, where users create item models, while the system aims to generate the variables. Nevertheless, the generation of the items from a knowledge model as in [11] requires that more complex knowledge is encoded in the data (e.g., what happens to water when the temperature decreases). The type and nature of data published as Linked Data need therefore to be further analyzed in order to support the development of such models for the fully automated generation items based on knowledge models.

We will focus our future work on the creation of an authoring interface for item models with the use of data sources from the Semantic Web, on the assessment of

item quality, on the creation of different types of assessment items from Linked Data sources, on the traceability of items created, including the path on the Semantic Web datasets which were used to generate the item, and on the improvement of data selection from semantic datasets.

Acknowledgments. This work was carried out in the scope of the iCase project on computer-based assessment. It has benefited from the TAO semantic platform for e-assessment (<https://www.tao.lu/>) which is jointly developed by the Tudor Research Centre and the University of Luxembourg, with the support of the Fonds National de la Recherche in Luxembourg, the DIPF (Bildungsforschung und Bildungsinformation), the Bundesministerium für Bildung und Forschung, the Luxembourgish ministry of higher education and research, as well as OECD.

References

1. Belluck, P. To Really Learn, Quit Studying and Take a Test. *New York Times*. January 20th, 2011
2. Karpicke, J. D., & Blunt, J. R. Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science*. (2011)
3. Gilbert, L., Gale, V., Warburton, B., & Wills, G. Report on Summative E-Assessment Quality (REAQ). Joint Information Systems Committee, Southampton. (2008)
4. Aldabe, I., Lopez de Lacalle, M., Maritxalar, M., Martinez, E., Uria, L. Arikurri: an Automatic Question Generator Based on Corpora and NLP techniques, ser. Lecture Notes in computer science, vol. 4053, pp. 584–594. Springer, Heidelberg (2006)
5. Lee, J. S. Y. Automatic correction of grammatical errors in non-native English text. PhD dissertation at The Massachusetts Institute of Technology. (2009)
6. Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation. *Knowledge Management & E-Learning: An International Journal (KM&EL)*, 2(3), 210. (2010)
7. Karamanis, N., Ha, L. A., & Mitkov, R. Generating multiple-choice test items from medical text: a pilot study. In *Proceedings of the Fourth International Natural Language Generation Conference*, pp. 111–113. (2006)
8. Lin, Y.C., Sung, L.C., and Chen, M.C. An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding. *Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007)*, pages 137-142. (2007)
9. Brown, J. C., Frishkoff, G. A., & Eskenazi, M. Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 819–826). (2005)
10. Sung, L.-C. Lin, Y.-C., Chen, M. C. The Design of Automatic Quiz Generation for Ubiquitous English E-Learning System. *Technology Enhanced Learning Conference (TELearn 2007)*, pp. 161-168, Jhongli, Taiwan. (2007)
11. Linnebank, F., Liem, J., and Bredeweg, B. Question generation and answering. *DynaLearn, EC FP7 STREP project 231526, Deliverable D3.3*. (2010)
12. Liu, B. SARAC: A Framework for Automatic Item Generation. In *2009 Ninth IEEE International Conference on Advanced Learning Technologies* (pp. 556-558).

Presented at the 2009 Ninth IEEE International Conference on Advanced Learning Technologies (ICALT), Riga, Latvia. (2009)

13. Xu, Y., Seneff, S. Speech-Based Interactive Games for Language Learning: Reading, Translation, and Question-Answering. *Computational Linguistics and Chinese Language Processing* Vol. 14, No. 2, pp. 133-160. (2009)
14. Lai, H., Alves, C., & Gierl, M. J. Using automatic item generation to address item demands for CAT. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. (2009)
15. Gierl, M.J., Zhou, J., Alves, C. Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *Journal of Technology, Learning, and Assessment*, 7(2). (2008)
16. Sarre, S., Foulonneau, M. Reusability in e-assessment: Towards a multifaceted approach for managing metadata of e-assessment resources. *Fifth International Conference on Internet and Web Applications and Services*. (2010)
17. Suchanek, F. M., Kasneci, G., & Weikum, G. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web* (pp. 697–706). (2007)
18. Mitkov, R., An Ha, L., & Karamanis, N. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(02), 177–194. (2006)
19. Foulonneau, Muriel, Cole, Timothy W. Strategies for reprocessing aggregated metadata. *European Conference on Digital Libraries. Lecture notes in computer science* 3652 , 290-301 (2005)
20. Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. A feasibility study of on-the-fly item generation in adaptive testing. *Educational Testing Service*. (2002)