EPiC Series in Language and Linguistics

Volume 1, 2016, Pages 361–371

CILC2016. 8th International
Conference on Corpus Linguistics

BiTeXting your food: Helping the gastro industry reach the global market

R. Rabadán[1], V. Colwell[2], and H. Sanjurjo-González[3]*

[1] University of León, León, Spain
rosa.rabadan@unileon.es
[2] University of León, León, Spain
vera.colwell@unileon.es
[3] University of León, León, Spain
hsang@unileon.es

Abstract

The technologization of cross-linguistic communication and the expansion of the learning of foreign languages has helped create new, non- linguist users. Corpus-based applications offer a way of responding to these new challenges. This presentation focuses on the design and building process of the BiTeX app, designed to write recipes in both En and Es through controlled language choices. The starting point is a custom-made, rhetorically and POS annotated, En-Es comparable corpus of recipes containing 135,912 words in the En and 145,449 in the Es subcorpus respectively. The BiText prototype has been developed using MongoDB [1], Express [2], Node.js [3] and jQuery [4], which allow for multiple concurrent connections to be handled without I/O blocks. BiTeX aims at helping improve international communication in the restaurant and catering community, as well as boosting collateral business niches in including recipe books, tourist-oriented websites, etc.

1 Introduction: What is Bitex?

BiTeX is a genre-specific prototype for a bilingual, bi-directional writing aid which uses a corpus-based controlled natural language (CNL), still in its developmental stages. BiTeX features the genre recipe and it is addressed to at least three target groups: i) gastro professionals, who are unfamiliar with the specific techniques that apply to professional writing, ii) educators involved in lifelong learning and continuing education, and iii) food bloggers and the like.

BiTeX helps produce recipes in both En and Es by providing guidance through controlled choices in writing in either language. It is not a list of dos and don'ts, nor is it a set of guidelines

*Co-author H. Sanjurjo-González is a doctoral student funded by the University of León. At the time of presenting this paper conducting his research at UCREL

[1] https://www.mongodb.org/ [Accessed 29.04.2016]
[2] http://expressjs.com/ [Accessed 29.04.2016]
[3] https://nodejs.org [Accessed 29.04.2016]
[4] https://jquery.com/ [Accessed 29.04.2016]

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 361–371

to be followed by the user. Rather it prompts users to draft their texts –in this case recipes- by making decisions on the basis of corpus-based information provided by the application.

Figure 1 illustrates the BiTeX task flow- The process starts from document structure recognition by the application of the genre-specific rhetorical information in one of the languages, i.e., the moves and steps that typically conform a recipe. Then the user is offered prototypical strings that include not only the necessary grammatical structures but also the range of lexical choices which are included in the bilingual dictionary(es). Next the application draws the correspondences between both languages to produce a genre-specific controlled language (CNL) that is currently under development. The result is a web-based application that helps produce bilingual recipes (Fig. 2).

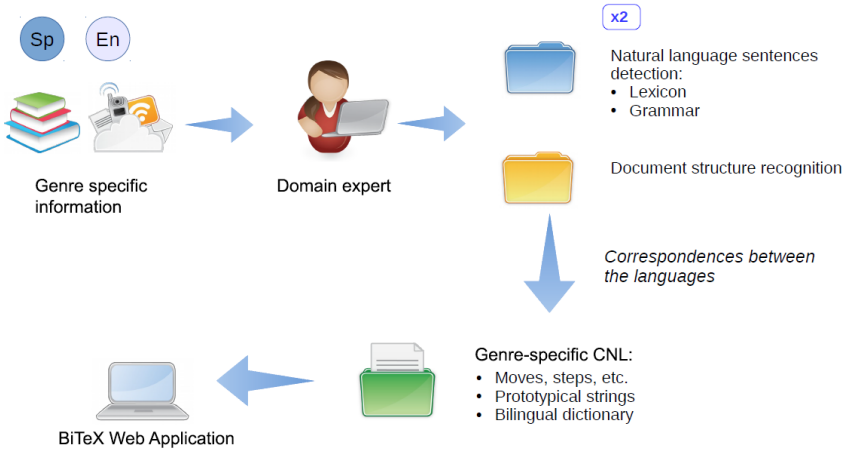

Figure 1: Application task flow

Nombre del plato:	**Name of dish:**
Macarrones con chorizo	Penne with chorizo
Ingredientes	**Ingredients**
400 gr. de macarrones	400 gr. penne
200 gr. de chorizo	200 gr. chorizo
800 gr. de tomate frito casero	800 gr. homemade fried tomato
Sal	Salt
Aceite de oliva virgen extra	Extra virgin olive oil
Preparación	**Method**
Poner cazuela ancha con agua hirviendo, añadir sal, macarrones y dejar cocer 20 minutos.	Place wide flat saucepan with boiling water add salt, Penne and boil for 20 minutes
Poner una sartén con aceite de oliva, chorizo y tomate frito y freír 4 minutos.	Heat skillet with extra virgin olive oil. Add chorizo, fried tomato and fry 4 minutes.
Escurrir los macarrones.	In a strainer drain the penne
Mezclar en una cazuela ancha el chorizo, tomate frito y los macarrones.	In a wide flat saucepan combine chorizo, fried tomato and penne.

Figure 2: BiTeX recipe

2 Background

BiTeX capitalizes on the concept of bi-text put forward by Brian Harris (1988). Originally a psycholinguistic concept defined as "a single text in two dimensions, each of which is a language (..) and co-exist in the translator's mind at the moment of translating" [7], in Translation Studies this concept has been seen as a way to partly explain concepts such as natural translation [6] or unit of translation [5]. In the field of corpus studies bi-text soon –and aptly- became a synonym for parallel concordances (original + translation) [13]. In this contribution bi-text goes a step further: it combines natural language processing capabilities with English ↔ Spanish parsing to (re)produce the two textual dimensions when prompted by human controlled writing in one of the languages. The design of a 'drafter language' [11] [4] enables users and the computer to produce the "two parts" of the same text simultaneously. In BiTeX directionality is not a variable: the human user can trigger the process starting from En or Es, the computer will always draw on a bi-recipe.

This 'drafter language' is a genre-restricted CNL where the grammatical and lexical choices are associated to particular rhetorical moves/steps. It has been derived from empirical data taken from real life language as represented in the BiTeX bilingual corpora.

3 What lies behind BiTeX

BiTeX drafter language draws on two bilingual corpora. C-BiTeX http://contraste2.unileon.es/web/es/corpus0_BiText.html is a custom-made bilingual comparable corpus specialized by genre- recipes- featuring materials from 2013 onwards sourced from food blogs. Concerning its size, C-BiTeX contains 135912 words in Es and 145449 in English distributed in 350 full texts in each of the languages. C-BiTeX has been annotated rhetorically using the ACTRES rhetorical tagger© http://contraste2.unileon.es/web/en/tools.html and also at POS level using Tree Tagger [12]. Linguistic information can be retrieved by means of the ACTRES browser©. P-BiTeX is a bi-directional parallel corpus, still under construction, containing dish names as quoted in C-BiTeX, their translations - L1 originals + L2 translations and viceversa- (e.g. *Cheesecake- Tarta de queso*) and a definition/description so as to overcome crosslinguistic misunderstandings and/or associations in this extremely culture bound area (e.g. *Gazpacho: A cold vegetable soup made of ripe tomatoes, pepper, onion, cucumber, a hint of garlic and bread crumbs mixed with olive oil, sherry vinegar and water*). P-BiTeX uses the MLCT bundle [10].

C-BiTeX has provided quantitative and qualitative empirical information concerning i) the rhetorical structure; i.e., moves and steps and their typicality range, ii) grammatical uses and prototypical strings and iii) genre/ move/ step-bound lexicography and phraseology in each of the languages En and Es. P-BiTeX is the source for opaque, culture-bound dish names, for which regular language pairing strategies do not offer an acceptable solution, e. g. *Welsh rarebit, olla podrida*, etc [5].

[5] A different situation is provided by innovative cuisine dishes (e.g. http://www.elbulli.com/catalogo/catalogo/index.php?lang=en), whose names are to be treated as 'new metaphors [9] when translating; i.e., to reproduce –when possible- the new association in the target language. Dish name generators such as Singer-Vine's [8] offer a festive approach to this problem.

4 Language work

The rhetorical structure is a corpus-based account of the genre 'recipe' which in both languages functions as an instructive text. If what it contains is clear in both languages, how this is conveyed is not necessarily so. Following [2], [1] and [3] twelve moves – two of them multimodal (photos, video)- have been identified for recipes. Of these, six have been found to be compulsory (figure 3).

MOVES - COMPULSORINESS			
M1 - Name of dish	100%	M7 - Option(s)	
M2 - Source and acknowledgements		M8 - Special dietary requirements (icons added)	100%
M3 - Background	100%	M9 – Storage and freezing	
M4- Technical info: type of dish, equipment, difficulty, time, etc.	100%	M10 -Notes	
M5 - Ingredients	100%	M11 - Photos	
M6 -Method	100%	M12 - Video tape	

Figure 3: Rhetorical structure of genre 'recipe'

Moves may include embedded steps that help organize the text. For example, M6 Method, which appears in every recipe in C-BiTeX, and is therefore compulsory in our structure, is composed of five different steps which are not compulsory with the exception of 6.1. Actions. Each step contributes distinct, relevant information:

M6. Method (100%)

- 6.1. Actions (100%)
- 6.2. Oven/ stove/ robot settings
- 6.3. Accompaniments
- 6.4 Serving suggestions
- 6.5 Wine pairing

Each of them is actualized by means of move-associated structures. Corpus-based extraction and the statistics provided by our tools have enabled us to identify a number of prototypical strings that contain invariable lexicogrammatical elements and variable choices according to need. These variable choices are accounted for in our controlled language by means of restricted move/ step dictionaries/glossaries. Prototypical strings are then a kind of controlled 'drafting lines' to be completed by the user when variation is marked in his/her working language. Different option types are indicated by different types of brackets, as in table 1.

Round brackets indicate that it is obligatory to fill in the gap and that the user must activate the corresponding restricted dictionary. Square brackets tell the user that the option is available, but not required. If selected, it is necessary to activate the corresponding restricted dictionary. Curly brackets signal that one of the choices offered in the string is required and prompt the user to select as appropriate. While round and square brackets normally indicate

Es	En
Servir caliente, acompañado de la salsa de chocolate en salsera aparte.	Serve piping hot with chocolate sauce on the side.
(ACCIÓN) [TEMPERATURA], {con/ acompañado de}[INGREDIENTE^N] [ELABORACIÓN][PRESENTACIÓN]	(ACTION) [TEMPERATURE] with [INGREDIENT^N][FARE] [PRESENTATION SUGGESTIONS]

Table 1: BiTeX controlled drafting lines

lexical information choice, curly brackets tend to mark functional, grammatical information. A superscript^N signals that the user must choose 'N' times, as necessary.

5 Relation between computing and linguistic work

In order to profitably use the rhetorical structure, controlled bi-strings and bilingual dictionaries mentioned above, they must be converted into a machine-understandable database. MongoDB was chosen mainly because it allows full stack JavaScript development. There are two MongoDB databases: one, named *structure*, contains the rhetorical structure and the prototypical strings (fig. 4). The other, named *bi-dictionary*, holds the bilingual restricted dictionaries (fig. 5)


```
{
    "_id":{"$oid":"567361f028342bb2f949c4ab"},
    "correspondencias":[{"en":"of","es":"de"},{"en":"with","es":"con"},{"en":"with","es":"a base de"}],
    "model_lines":
        {
            "0":
            {
                "en":"(ELABORACION) {de/con/a base de} (ESTADO) |INGREDIENTE 1| {de/con} |INGREDIENTE 2|",
                "en_ej":"Pie made of baked pilchards with eggs and potatoes, covered with a pastry crust",
                "es":"(ELABORACION) {de/con/a base de} |INGREDIENTE 1| (ESTADO) {de/con} |INGREDIENTE 2|",
                "es_ej":"Gazpacho-\u003e Sopa fría de tomate, pimiento, cebolla, pepino, ajo y miga de
                        pan emulsionada con agua, aceite de oliva y vinagre de Jerez"

            }
        },
    "nivel":0.0,
    "orden":1.0,
    "src":"glyphicon glyphicon-list-alt",
    "tipo":"basico",
    "titulo":
        {
            "en":"Description",
            "es":"Descripción"
        }
}
```


Figure 4: Sample of MongoDB *structure* file.

6 BiTeX computing work

BiTeX features a three-tier structure: data, logic and views. Fig. 6 displays the web application architecture and its flow.

```
{
    "_id":{"$oid":"56735f93d1a03efc4ca0a67a"},
    "categoria_en":"(ACTION)",
    "categoria_es":"(ACCION)",
    "en":["soften"],
    "es":["ablandar "]
}
{
    "_id":{"$oid":"56735f93d1a03efc4ca0a67b"},
    "categoria_en":"(ACTION)",
    "categoria_es":"(ACCION)",
    "en":["place "],
    "es":[" introduce"]
}
```


Figure 5: Sample of MongoDB *bi-dictionary* file.

- Data: As described above, MongoDB is the machine-understandable format that employed in the BiTeX application. The data component consists of the 'bi-dictionary', 'structure' and multimedia files as configuration files.
- Logic: BiTeX was developed using Node.js environment which provides a fast and efficient architecture for data-intensive treatment for real time applications. Node.js is responsible for data conversions, data loading and data saving as well as hosting the application by means of a Web application framework called Express.js. Therefore, Node.js ensures the correct operation of the entire process.
- Views: Lastly, the interface views were done mainly by means of HTML5 [6] and Bootstrap framework [7]. jQuery is also employed for some visual effects and data treatment in the user interface. Thus, BiTeX boasts a responsive, multi-platform, user-centered and user-friendly interface. Because BiTeX is a Web application it can be accessed via a Web browser and is installation-free.

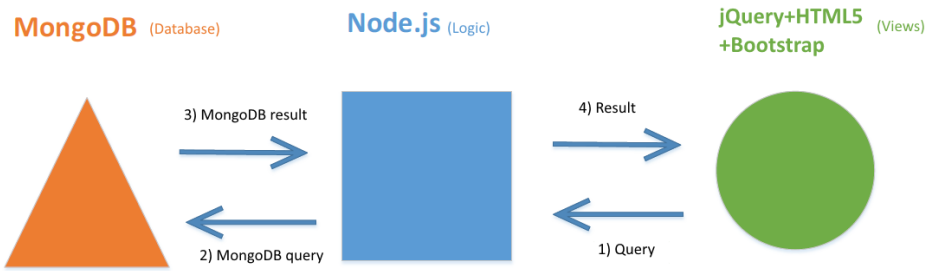

Figure 6: Web Application architecture and prototypical flow.

[6] https://www.w3.org/TR/html5/ [Accessed 29.04.2016]
[7] http://getbootstrap.com/ [Accessed 29.04.2016]

7 How does the BiTeX app work?

The BiTeX bi-writing application uses the *bi-dictionary* and *structure* databases to create a custom-made user interface. First, the basic, default user interface is shown. Next, the *'structure'* database is accessed and text-input sections are provided for each new constituent part. These new sections typically have a title section, a set of 'model drafting strings' together with the appropriate corpus-based examples and an interactive area, where the user will 'ask' the application to activate the *bi-dictionary* database or to deactivate the brackets. Then, these controlled choices will guide the user in the writing of a recipe in En or Es. BiTeX also offers other interactive fields that require the user to upload static images, videos or other miscellaneous information.

But, how is this interaction realised and how is the translation of the controlled choices accomplished? As shown in figure 7, the user has to select one of the proposed 'model drafting strings' which is then inserted in the input text area. Model drafting strings are actually *controlled bi-lines* that are formed by two lines, one in En and one in Es (see Harris's quote in section 2 above). Depending on which language is selected, the other language is hidden from the user's view.

To complete a model drafting string, the user must activate the restricted bilingual dictionary. This means that a query is executed in the *bi-dictionary* database. A query searches for correspondences between the En and/or Es items and their context. Context here means the exact rhetorical section, and selected model string where the query is launched. A clear example of how a query is executed is illustrated in figure 8.

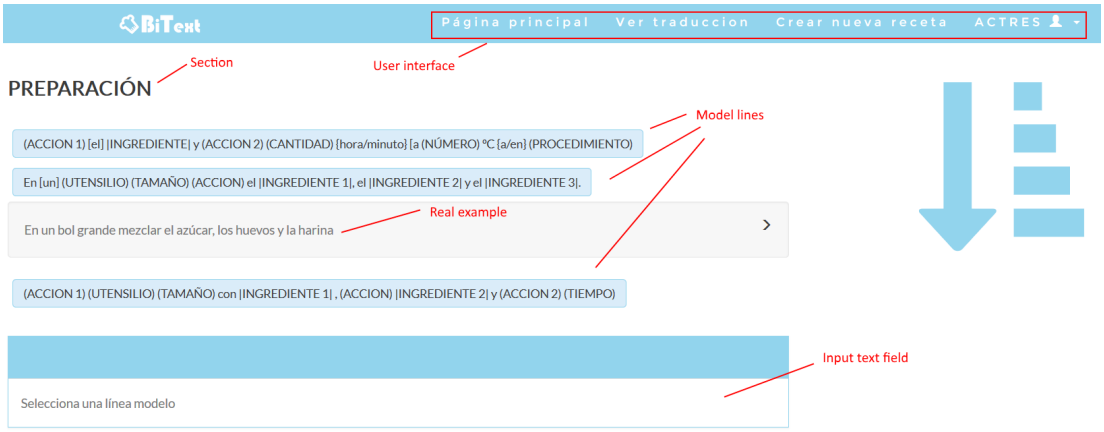

Figure 7: Controlled BiTeX choices for Es speakers

8 Using BiTeX

In order to explain how to use BiTeX properly we will fill out one section starting from Es, the "Preparación" (preparation) section. Figure 9 shows the BiTeX user interface described in previous section 7 above. An interactive area appears below the model lines or strings and the examples. This is the input area where the user will draft his/her text.

If we click on the second row:

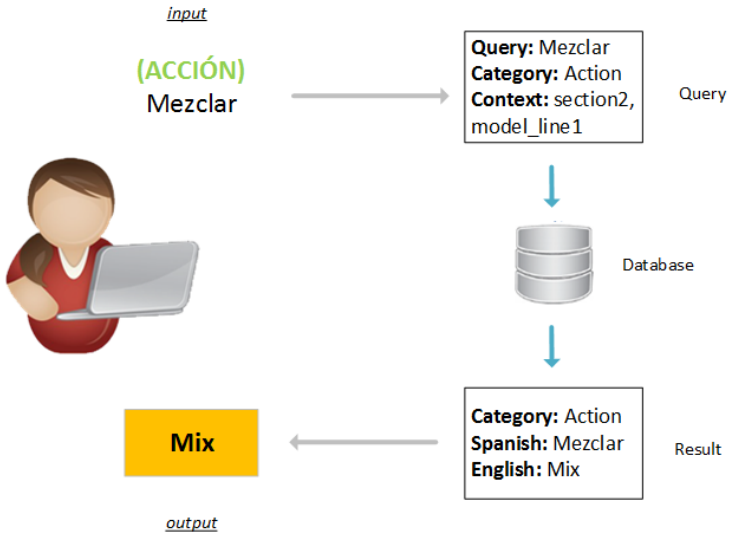

Figure 8: Example of query in BiTeX

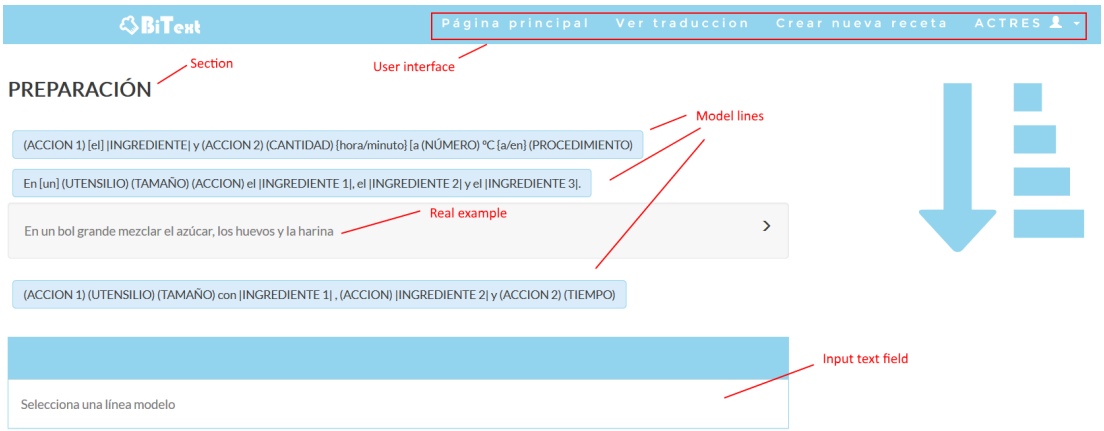

Figure 9: 'PREPARACIÓN' section

"En [un] (UTENSILIO) (TAMAÑO) (ACCIÓN) el |INGREDIENTE 1|, el |INGREDIENTE 2| Y EL |INGREDIENTE 3|"

A real example pops up to provide guidance to the user:

"En un bol mezclar el azúcar, los huevos y la harina"

Next, we click on the arrow to insert the model line/ string to be used in the input text field as shown in Figure 10.

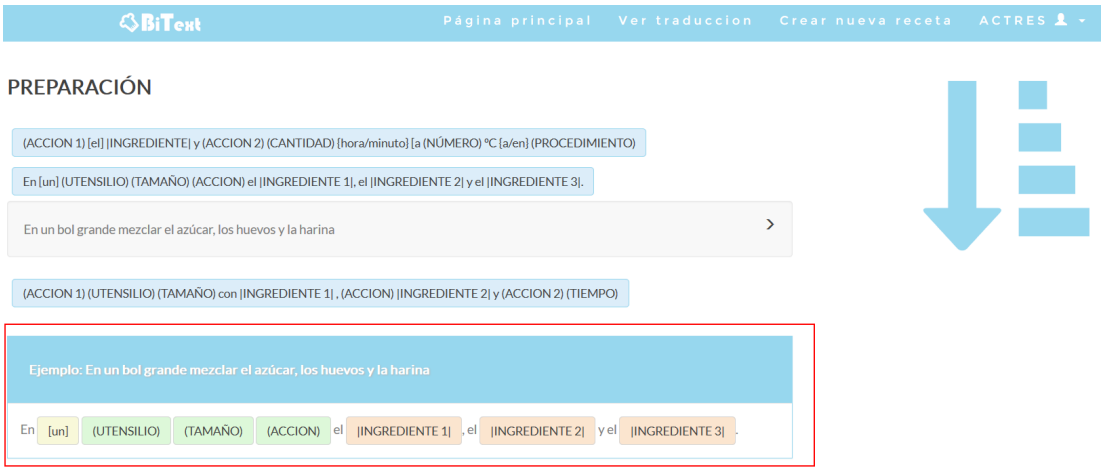

Figure 10: Inserting a model line/ string in input text field

Once the model string has been selected and inserted in the interactive area, the user goes on to make his/her selections, which are color-coded, as shown in figure 11. The first element in the string is yellow-coded "[un]", which indicates optional content, as described in section 4 above. A dialog area will pop up asking whether the element is to be kept or deleted.

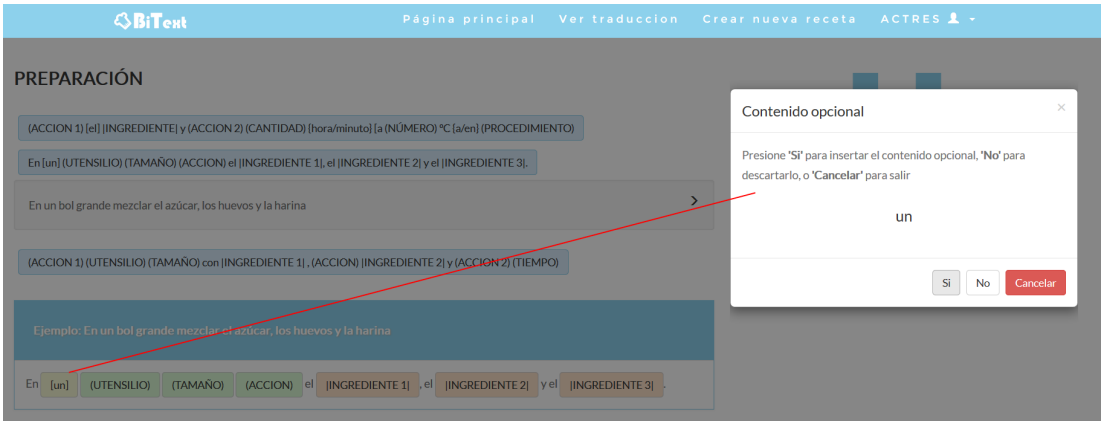

Figure 11: Selecting optional content in string example

If the user clicks on green-coded "(UTENSILIO)" a new dialog area will pop up. In this case he/she is prompted to search for an appropriate item to fill in the slot. In this case, the choice is "bol".

If we repeat the process until the string is completed and click on "Ver traducción" in the top bar of the user interface, the En string appears, as illustrated in figure 13.

By replicating this process throughout all the remaining sections, a complete and correct English translation of the desired recipe is attained.

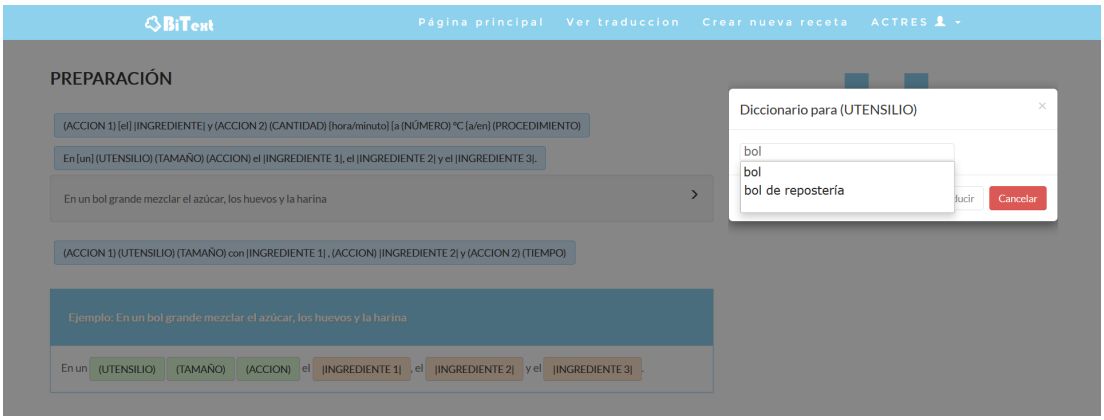

Figure 12: Selecting compulsory content in string example

Figure 13: Completed Es-En bi-string

9 Conclusions

In the light of ever-increasing market demands for English as the international business language, BiTeX is one response to the linguistic needs of the Spanish L1 gastro professionals. Results from informants in the proofs of concept (PoC) indicate that this is not only an innovative application that meets their needs but that it does so in ways that other resources do not.

While users report on the attractive and efficient aspects of BiTeX, the fact that it is not fully automatic is the one aspect they would consider a weakness. For the app builder, however, BiTeX is not only time consuming and in need of extensive corpus-based contrast, it also requires extensive testing among different user groups.

Nevertheless in its present form BiTeX has clearly identifiable strengths. For the builder, given its corpus-based nature, it offers a fully natural drafter language, which is customizable and expandable as well as innovative responsive design technology. Ongoing work is focused on

making BiTeX trial drafter language more robust and its underlying system replicable in other domains. For the user it is a friendly and time-saving application which requires no training and, furthermore, is dependable in terms of language correction and acceptability.

References

[1] Vijay Bhatia. *Worlds of written discourse: A genre-based view*. A&C Black, 2004.

[2] Vijay Kumar Bhatia. *Analysing genre: Language use in professional settings*. London: Longman, 1993.

[3] Douglas Biber, Ulla Connor, and Thomas A Upton. *Discourse on the move: Using corpus analysis to describe discourse structure*, volume 28. John Benjamins Publishing, 2007.

[4] Antonis Bikakis, Paul Fodor, and Dumitru Roman. *Rules on the Web: From Theory to Applications: 8th International Symposium, RuleML 2014, Co-located with the 21st European Conference on Artificial Intelligence, ECAI 2014, Prague, Czech Republic, August 18-20, 2014, Proceedings*, volume 8620. Springer, 2014.

[5] Toury Gideon. Descriptive translation studies and beyond. *Amsterdam & Philadelphia: John Benjamins*, 1995.

[6] Brian Harris. *The importance of natural translation*. University of Ottawa, School of Translators and Interpreters, 1977.

[7] Brian Harris. Bi-text, a new concept in translation theory. *Language Monthly*, 54:8–10, 1988.

[8] Jeremy. Strawberries with sambuca baba and acid honey. the el bulli dish-name generator. http://www.slate.com/articles/life/food/2011/07/strawberries_with_sambuca_baba_and_acid_honey.html?from=rss, 2011. Accessed 27.04.2016.

[9] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago, 1980.

[10] Scott Piao. Mlct (multilingual corpus toolkit). https://sites.google.com/site/scottpiaosite/software/mlct, 2016.

[11] Richard Power and Donia Scott. Multilingual authoring using feedback texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1053–1059. Association for Computational Linguistics, 1998.

[12] Helmut Schmid. Treetagger - a language independent part-of-speech tagger. http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html, 2013. Accessed 27.04.2016.

[13] Jörg Tiedemann. Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165, 2011.