Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

How can Writing Tasks be Characterized in a way serving Pedagogical Goals and Automatic Analysis Needs?

Martí Quixal and Detmar Meurers

Abstract

The paper tackles a central question in the field of Intelligent Computer-Assisted Language Learning (ICALL): How can language learning tasks be conceptualized and made explicit in a way that supports the pedagogical goals of current Foreign Language Teaching and Learning and at the same time provides an explicit characterization of the Natural Language Processing (NLP) requirements for providing feedback to learners completing those tasks? We argue that the successful implementation of language learning tasks that can be automatically assessed demands a design process in which both pedagogical and computational requirements feed each other.

Extending well-established work in Task-Based Instruction (TBI) and Language Testing, we propose a framework that helps us (a) elucidate the formal features of foreign language learning activities, (b) characterize the relation between the expected and the actually elicited learner language, and (c) assess how the variability in the learner responses impacts interpretation and computational techniques for the automatic analysis of learner language.

To validate our approach we apply the framework to two writing tasks for learners of English as a foreign language. Our analysis highlights the relevance of spelling out the pedagogical and linguistic goals of language learning tasks in order to successfully characterize the language expected and actually elicited in learner responses, to evaluate the realization of the task, and to support the design of effective NLP strategies. Given the combination of design- and data-driven perspectives, the framework supports an iterative approach to the creation of language learning tasks and ICALL materials.

KEYWORDS: ICALL TASK DESIGN; TASK-BASED CALL; AUTOMATIC ASSESSMENT

Affiliation

Seminar für Sprachwissenschaft and LEAD Graduate School Eberhard-Karls-Universität Tübingen.
email: marti.quixal@gmail.com (corresponding author)

CALICO JOURNAL VOL 33.1 2016 19–48

doi : 10.1558/cj.v33i1.26543

equinox www.equinoxpub.com

equinoxonline

1. Introduction and motivation

Since the late 1970s, researchers in Natural Language Processing (NLP), Foreign Language Teaching and Learning (FLTL), and Second Language Acquisition (SLA) have worked on the design, development, and integration of computer-delivered language learning materials supporting the automatic evaluation of language-mediated responses (Heift & Schulze, 2007; Weischedel & James, 1978). While there have been some successful projects coupling pedagogical needs with computational capabilities (cf. Schulze, 2010), the dialog between FLTL and NLP has been full of controversy (Borin, 2002; Heift & Schulze, 2007; Salaberry, 1996). The lack of state-of-the-art NLP analyses integrated in CALL is as frequently criticized as the lack of pedagogically motivated activities and the perceived unreliability of the NLP. This has in general prevented CALL systems integrating NLP from being perceived as relevant for Communicative Language Teaching (CLT) (Heift & Schulze, 2007: 1, 221).

At the same time, connecting NLP to CLT is crucial given its widespread adoption both as a research and as an instruction paradigm (Brown 2007: p. 241; Ellis, 2003: Ch. 1; Richards & Rodgers, 2001). As argued in Amaral & Meurers (2011: 9–11), for systems to be compatible with CLT approaches, they must be able to process meaning (also referred to as *content*) as well as form. In addition, for current language technology to be effective and efficient, constraining the language to be elicited from learners is absolutely central (Amaral & Meurers, 2011: 10). It is not generally possible to reliably interpret learner language without taking the nature of the task and the learner into account (Meurers, 2015). Designing tasks that are pedagogically meaningful and that can be reliably assessed using NLP strategies thereby constitutes the fundamental challenge of Intelligent Computer-Assisted Language Learning (ICALL).

In this paper, we aim at making explicit the relationship between a pedagogically principled design of Foreign Language (FL) learning activities and the characteristics of NLP-based feedback generation. We propose a methodology for the design, implementation and evaluation of ICALL materials that leverages insights from Task-Based Instruction (TBI) and Language Testing. This makes it possible to relate task characteristics, the linguistic properties of the learner language they elicit, and the processing requirements they entail.

This paper consists of four further sections. Section 2 argues for the use of tasks to connect pedagogical needs and computational requirements and introduces the FLTL and SLA work that inspires our approach to the design of ICALL tasks. Section 3 introduces our approach and exemplifies its application with two English as a FL (EFL) writing tasks developed for a language learning course. Section 4 analyzes the response variation in two small sets of

learner responses to the tasks presented in order to illustrate their relevance for the development of ICALL materials and discusses the pedagogical and computational aspects that can be derived from the analysis. Finally, Section 5 concludes and sketches avenues for future research.

2. Tasks as the connecting point between FLTL and NLP

As argued in Bailey and Meurers (2008: 107), the development of processing strategies supporting the generation of pedagogically meaningful feedback requires an understanding of the connection between FL learning activities and the linguistic variation expected for the learner responses they elicit. In the present paper, we want to argue that this connection should be made fully explicit to support both the validation of the learning materials and the implementation of automatic assessment strategies. We propose to use tasks, as in Task-Based Instruction, as a key instrument to support this connection and around the concept of task we present a framework that characterizes the relationship between a task's pedagogical objectives and the language produced by the learners.

TBI is a widespread method of syllabus design in which units of work are conceived according to the goals and the needs of a specific task that learners ought to be able to do in their everyday lives (Brown, 2007: 242). The definition of task is not free of controversy (Brown, 2007: 242–243; Ellis, 2003: Ch. 1; Nunan, 2004: Ch. 1). To make things concrete, we here follow Nunan's definition of a task:

> [A] piece of classroom work that involves learners in comprehending, manipulating, producing or interacting in the target language while their attention is focused on mobilizing their grammatical knowledge in order to express meaning, and in which the intention is to convey meaning rather than to manipulate form. (Nunan 2004: 10)

Such a pedagogical task includes the instructions and the means given to learners for practicing the targeted skills, which are related to particular linguistic structures to be elicited (Ellis, 2003: 16–17). To qualify as a communicative task, a task must be focused on meaning and have a communicative outcome, allowing students to choose their linguistic structures within a semantically delimited space. A task should also involve cognitive processes and communicative strategies actually used in real-life interactions (Ellis, 2003: 9–10).

2.1. Impact of task features on learner language

Researchers and practitioners in SLA and FLTL characterize tasks from the perspective of their pedagogical goals and the targeted linguistic structures so that they can: (a) determine which tasks are useful for what cognitive and communicative features; and (b) produce a balanced repertoire of linguistic goals to be worked on when designing course materials.

Bachman and Palmer (1996: Ch. 3) offer a fine-grained framework for the macro-level characterization of task-based tests. It includes aspects such as the *physical setting*, the *rubric* (the structure of the test), the *input* given to the learner, the *expected response* (learner production), and the *relationship between input and response*. As Bachman and Palmer point out, this framework conceived for test design can in principle be applied to any kind of learning material that involves the evaluation of the language elicited from learners.

2.2. Characterization of the response language

When bringing TBI into practice, Estaire and Zanón (1994) and Willis (1996) spell out in detail how instructors can anticipate what a given task elicits from learners. Estaire and Zanón (1994: 30, 58) propose a framework to design task-based curricula involving a task development cycle that (a) integrates content, objectives, methodology, and evaluation, and (b) encourages teachers to work backward from the materials to identify contents, procedures, and tools for assessment.

In their framework, Estaire and Zanón (1994: 29) suggest that material designers spell out the linguistic contents of learner responses on the basis of the thematic content. Such specifications should include the functions, the linguistic features and the communicative features that learners need to learn and develop further for a given task. They suggest that specifications can be done *a priori*, if the task is closed enough. However, if the task is more open or free, the linguistic features of the responses can be created retrospectively to evaluate the achievement of the pedagogical goals.

By combining Estaire and Zanón's syllabus design framework with Bachman and Palmer's notion of relationship between input and response, we obtain the linguistic characterization of learner responses not only as a function of its thematic content, but also as a function of the resources that learners are given to complete the task. In particular, we aim at determining and exploiting what Bachman and Palmer (1996: 52–54) define as the *scope* and the *directness* of the relationship between input and response. Scope ranges from narrow to broad and relates the amount of input to be processed by learners with the amount of text that they are expected to produce. Directness, on the other hand, refers to the degree to which the expected response can be based on information provided with the input, compared to needing to rely on world knowledge or context. This perspective is worked out in the Response Interpretation Framework, a framework component that we spell out in the next section.

The next section presents a framework that takes advantage of a detailed pedagogical characterization of tasks to inform the automatic assessment needs. Such a characterization includes details on the target linguistic items,

the correctness of the responses, the grading rubrics, the expected learner products, the type of assessment, and the interrelationship between linguistic analysis and the generation of feedback.

3. Addressing the need for an explicit analysis of tasks

3.1. Components of our approach

Our methodology for a pedagogically and computationally principled design of task-based ICALL activities has three components: the *Task Analysis Framework (TAF)*, *the Response Interpretation Framework (RIF)*, and the *Automatic Assessment Specification Framework (AASF)*. The methodology, integrating and extending distinctions from SLA (Ellis, 2003; Littlewood, 2004), FLTL (Estaire & Zanón, 1994), and Language Testing (Bachman & Palmer, 1996; Douglas, 2000), is envisaged to cover the elements needed to link task characteristics as discussed in language teaching with the specification of analysis techniques in computational linguistics.

The TAF supports the characterization of the pedagogical features of tasks at a macro level. Learning and communicative goals will determine, among others, whether a constructed or a production response is more effective to implement the materials in the classroom. The TAF embraces the characterization of language learning materials in general, not just CALL tasks.

The second component, the RIF, facilitates a detailed characterization of the linguistic structures that a task is expected to elicit from learners, and it focuses on production responses. It provides essential information on the freedom of learners to choose the linguistic resources to complete the task. It includes a formal specification of the expected thematic and linguistic contents, and the task's criteria for correctness – all of them key to specify the needs of the NLP-based feedback generation functionality and the pedagogical evaluation of learner responses. The RIF also facilitates the generation of sample responses that guide the design and evaluation of NLP-based assessment strategies.

Finally, the AASF takes the information provided by the two previous components and provides detailed specifications for the implementation of the NLP module and the feedback generation logic. To support a modular architecture, it distinguishes a language analysis module and a feedback generation module.

3.1.1. EFL tasks for a task-based course

Throughout the article, we use two English L2 tasks to exemplify our framework. These were developed as part of the ALLES (Advanced Long-distance Language Education System) project. This was an EU-funded project aiming to develop CALL materials following a TBI approach, including NLP-based

individualized feedback for both formative and summative assessment (Badia *et al.*, 2004; Schmidt *et al.*, 2004). The materials were designed for the instruction of Language for Specific Purposes targeting B2 and C1 CEFR-level learners of Catalan, English, German, and Spanish in the domain of business and finance.

The two ALLES activities that we use are two of the writing tasks developed for learners of English. The first is the *customer satisfaction questionnaire task*, which requires limited production responses. The second one, we refer to as the *course registration task* and requires an extended production response. The customer questionnaire task is used to exemplify the first two components, the TAF and the RIF, while the course registration task is used to exemplify the AASF. Authentic learner responses to both tasks are used in Section 4 to empirically validate the usefulness of our task development framework for the purposes of response evaluation and annotation and to inform the development of NLP-based feedback generation strategies.

3.2. Task Analysis Framework (TAF)

The TAF consists of eight rubrics: (1) the *Description*, an informal characterization of the task; (2) the *Focus*, to specify a focus on form, meaning, or both; (3) the *Outcome* product; (4) the cognitive *Processes* that the designer aims at practicing; (5) the *Input* data given to the learner to complete the task; (6) *Response type*, constructed, limited or extended response; (7) the *Teaching goal*, one of Littlewood's categories (2004: 322), and finally (8) the *Type of Assessment* required.

Table 1 illustrates the application of the TAF to a language learning task from the ALLES materials. The task *Stanley Broadband customer satisfaction questionnaire* was part of the lesson *Customer Satisfaction and International Communication*.

Table 1: Example TAF analysis

Description Writing the questions for a customer satisfaction questionnaire for the product *Stanley Broadband service*
Focus Meaning and form
Outcome Five interrogative sentences for the questionnaire
Processes Ask people about their opinion and intention with respect to a product or service; use of relevant linguistic structures to ask for information
Input Five items, each requiring a separate response and including a hint on the topic and the linguistic structures
Response type Limited production response: a sentence per item
Teaching goal Communicative language practice
Assessment Formative

It requires short responses whose contents will be constrained by the fictional professional setting of the task and by the input data provided. While other aspects of this task will be discussed below, the TAF highlights that it involves a language-mediated response requiring formative assessment and that the task is characterized as communicative language practice.

3.3. Response Interpretation Framework (RIF)

The RIF is only applied to tasks with language-mediated responses. It characterizes the nature of the expected response: the thematic and linguistic contents, also known as topical and linguistic knowledge (Bachman and Palmer, 1996: 54). It consists of six rubrics:

(1) **Instructions:** The *language* of the instructions, the *channel* of presentation, and the *specification of the procedures* (Bachman & Palmer, 1996: 50–51).
(2) **Input:** This rubric includes the *input data*: the material to be processed, its language properties and the *prompt* used to set up a communicative situation (Douglas, 2000).
(3) **Expected response:** The *format* (oral or written, lengthy or short) and the properties of the expected *language*, including the analysis of the *relationship between the input and the response* in terms of *scope* and *directness* (Bachman & Palmer, 1996: 54; Ellis, 2003: 289–291, 312).
(4) **Thematic content:** This rubric characterizes the information to be included in the response in terms of *entities* and *relations* (borrowing NLP terminology used to refer to people, objects, or concepts and the relations among them).
(5) **Linguistic content:** The linguistic properties of the expected responses regarding text structure and rhetorical organization, functional, grammatical and lexical contents (Estaire & Zanón, 1994: 30, 58).
(6) **Assessment criteria:** Specifications for the evaluation of learner output including the *criteria for correctness* and a *scoring procedure* (Bachman & Palmer, 1996: 52).

3.3.1. Exemplifying the RIF

Instruction, input data, and the relationship between input and response. Table 2 analyzes the first three features of the first and the last item in the customer satisfaction questionnaire task: prompt, instructions, and input data presented to the learner. The prompt of this task requires the learner to adopt the role of an employee that has to produce a customer satisfaction questionnaire.

Table 2: RIF example: instructions, input data and relationship between input and response

PROMPT	Imagine you work for Stanley Broadband. You have just listened to the interviews with Trevor and Janet. You would like to improve the service that Stanley Broadband offers. You have to compose a questionnaire to find out more about how to improve your company's service.
INSTRUCTIONS	Your task is to use the clues given for each box and write the necessary question.
EXAMPLE	0. Ask what customers thought about the cost of Stanley Broadband compared to other companies who provide Internet services. Include the words *Did you . . . find . . . expensive . . .*? *Did you find Stanley broadband more expensive than other broadband service providers?*
INPUT DATA	1. Ask about customers level of satisfaction with Stanley Broadband. Write a question beginning with *How. . . .?*
RESPONSE	
	(. . .)
INPUT DATA	5. Ask customers to describe future improvements they would like to see in the Stanley Broadband service. Begin the question with *What improvements . . .*?
RESPONSE	

Prompt, instructions and input data are in English, and the channel is textual. The instructions include an example that shows learners the expected production procedure.

Learners are required to use a set of clues given to produce five interrogative sentences. Each item includes input data that follows the pattern 'Ask about X' and 'Use the expression and/or word Y in your answer'. The task exemplifies a narrow relationship between input and response. The information to be processed is short, and so is the required response. All items present a direct relationship between input and response: both topical knowledge and language knowledge are restricted by the input data.

Thematic and linguistic content. Table 3 exemplifies the next two RIF features of the task, thematic and linguistic content, for the first item (Estaire & Zanón, 1994: 29). The table spells out the details of the thematic content, and the linguistic contents. The former include specific entities: (the product, the interviewer, the customer) and relations (the interviewee having an experience with and an opinion about a product). The latter specifies the functions,

specific exponents for the functions, specific syntactic and lexical resources and, in this case, specific pragmatic and orthography requirements.

Table 3: RIF example: thematic and linguistic contents

Thematic content of the expected response	
Entities	– Stanley Broadband – Your interviewee (your customer) – The interviewer
Relations	– Interviewee has an opinion or an experience as user of Stanley Broadband – Interviewer asks about the the level of satisfaction of interviewee using Stanley Broadband
Linguistic content of the expected response	
Functional	– Ask customers about satisfaction with a product *How satisfied are you with …* *How happy are you with …* *How much did you like using …*
Syntactic	– Use word order of wh-questions
Lexical	– *how, Stanley Broadband, you, satisfied, happy, satisfaction, …*
Pragmatics	– Use the appropriate register. – Use an interrogative sentence beginning with *how*
Orthography	– Use the appropriate spelling.

Assessment criteria. The next step is to spell out the Assessment criteria for a task that requires formative assessment. First, it requires a question: (a) asking for the level of satisfaction of the addressee, (b) including a reference to the Stanley Broadband service; and (c) starting with *how*. Second, the task requires language showing the appropriate use of: (a) word order in interrogative sentences; (b) register; and (c) spelling. On this basis, a set of potential correct responses can be generated to constitute a gold standard for the analysis. The term *gold standard* is used here as in NLP to refer to a set of analyzed linguistic elements against which a particular system should be compared to prove its validity. Example (1) shows two of the possible correct responses, and (2) exemplifies further options for conveying the intended meaning, using synonymy and a different exponent of function.

(1) a. How satisfied are you with the Stanley Broadband service?
    b. How satisfied are you with Stanley Broadband?

(2) a. How happy are you with Stanley Broadband?
    b. How much did you enjoy using the Stanley Broadband service?

The number of correct responses given a task and the grammar of the language being used in principle is potentially infinite (Nagata, 2009: 563–564). At the same time, when using a given task with students, the analysis of the subset of responses actually used by a given learner population can inform the linguistic and feedback generation modules. This allows for a systematization of the most frequent correct and incorrect responses, and common error types, using a corpus-based approach.

3.4. Automatic Assessment Specification Framework (AASF)

The AASF as the third component is needed to support the design and development of the NLP-based assessment functionality on the basis of the task and response characteristics spelled out in the TAF and the RIF. The AASF consists of the *Specifications for Automatic Linguistic Analysis (SALA)* spelling out an NLP-oriented specification for the analysis of the learner responses, and the *Specifications for the Feedback Generation Logic (SFGL)* making explicit the link between the analysis and the feedback to be generated. While the SALA and the SFGL are intended as specifications for the automatic analysis, they crucially integrate SLA and FLTL elements, from the nature of the linguistic constructs that are relevant to identify aspects of complexity, accuracy, and fluency to those needed for the provision of task-specific individualized feedback (Housen & Kuiken, 2009). Spelling out the task features that we present in the following is an endeavor to be performed in cooperation between FLTL/SLA experts and NLP developers since it requires the combination of experience in: (a) material design and learner language development; and (b) in the development of software for the automatic analysis of language.

The automatic analysis software and feedback generation logic are of a relatively technical nature, which in the actual project context were made accessible to teachers through a dedicated authoring system. While this is spelled out in detail in Quixal (2012: Ch. 11), for space reasons we here limit ourselves to discussing the conceptual side that made it possible to create the authoring system as an interface between the FLT/SLA and the NLP world.

We first exemplify the AASF based on the course registration task, before briefly sketching the implementation of the automatic assessment for this task in Section 3.4. In the course registration task, learners are expected to engage in a role play activity that requires them to write an email to the Human Resources department to register for a course. To do so, they are given a series of input material including a list of courses, a calendar page and voice mail from their manager. We here focus on the AASF – the full TAF and RIF analysis of this task can be found in Quixal (2012: Ch. 7).

3.4.1. Exemplifying the SALA

Table 4 shows the specifications for automatic linguistic analysis derived from the thematic content of the underlying RIF analysis for the course registration task, which requires an extended production response, and formative and summative assessment. The first row in the RELATIONS section specifies how to evaluate whether the learner stated the department in which (s)he is working, e.g., a relation between the email author and the department name (*NE:DeptName*). The '(…)' indicate where for space reasons the list is not completely spelled out.

Table 5 shows the specifications derived from the linguistic contents. The levels of linguistic description contained will depend on the activity's design and on the pedagogical needs, rather than on a complete characterization of each and every linguistic level.

Table 4: Example SALA: thematic contents of the RIF

Reference	Textual cue	Code	Module
Entities			
The email author	- *I* - *(My name is) X* - *(Sincerely,) X*	NE:EmailAuthor	IE (ent.)
Department's name	- *Marketing Department*	NE:DeptName	IE (ent.)
Course names	- *Business Communication* - *E-commerce & E-business*	NE:CourseName	IE (ent.) (…)
Relations			
State your department	- *I work for the Department of X* - *I am in the Dept. of X*	Rel:YourDepartment+ NE:DeptName	IE (rel.)
Tell what course you want to attend to	- *I am interested in the course on X* (…)	Rel:CourseToTake	IE (rel.)
Tell why this course can benefit you in the future	- *this course will be useful for the projects I will be involved in the future* (…)	Rel:UsefulFutProjects	IE (rel.) (…)

Tables 4 and 5 reflect the different types of processing strategies required: From Information Extraction (IE) modules to specific lexical or morphological analyzers, through spell and grammar checkers or shallow discourse parsers.

Table 6 then specifies linguistic complexity measures computed based on the learner responses. As before, what is spelled out here is motivated by the pedagogical approach and the objectives of the task.

3.4.2. Exemplifying the SFGL

The specification of the feedback generation logic for the summative feedback requires the definition of how the feedback messages are to be generated based

Table 5: Example SALA: linguistic contents of the RIF

Reference	Textual cue	Code	Module
Functional content			
Expressing interest	- *I am interested in* X	Func:ExpressInterest	IE (exp.)
Introduce oneself	- *my name is* X, *I am X* (. . .)	Func:IntrodOneself	IE (exp.)
(. . .)			
Syntactic content			
Describe/Report with Present Simple	- *I* work *for* X - *My boss* recommends *me* X	Syn:PresentTense	Morph. analysis (. . .)
Lexical content			
Domain-specific vocabulary	- register, apply, course, schedule, permission, manager, authorisation (. . .)	Lex:DomainVocab	Lexicon
Pragmatic content			
Greeting expression	- *Dear Sir or Madam* - *Dear colleagues,* - *To whom it may concern*	Prag:Greeting	IE (prag.)
Complimentary close	- *Sincerely yours*, . . .	Prag:ComplClose	IE (prag.) (. . .)
Orthography			
Appropriate spelling	- Word spelling and punctuation	SpellingOk	Spell checker

on the interpretation of the learner language resulting from the automatic linguistic analysis. Table 7 illustrates the feedback generation strategies envisaged for the different student performances.

Table 7 uses the information defined in Tables 5 and 6 to generate grades and messages. For instance, taking the communicative contents as an example, the very first row states that if all the expected thematic content (TC) and linguistic content (LC) items are identified (i.e., $TC = 6$ and ($\wedge$) $LC = 4$) then the response receives grade 4 and the feedback 'Very good. You use the expected functions adequately.'

3.5. Implementation of the automatic assessment module

To sketch the implementation side of the automatic assessment for the course registration task for which we saw the SALA spelled out in Table 4, let us zoom into the thematic content required for *Tell what course you want to attend to* included in that table. Figure 1 shows the rule-based linguistic patterns that were actually used to recognize the fragment of the response for the thematic content *Rel:CourseToTake*.

Table 6: Example SALA: linguistic complexity measures

Reference	Textual cue	Code	Module
Lexical contents			
Word fluency	- Length in words	No. of words	Tokenisation, lexicon
Specific vocabulary	- Domain specific terms	Specific vocabulary	Domain lexicon
Sentence structure and accuracy			
Syntactic complexity	- Simple and complex sentences	No. of sentences	Sentence segmentation & synt. analysis
	- Presence of discourse markers	No. of discourse markers	POS tagging and discmarker lexicon
Accuracy	- Grammar errors	No. of grammar errors	Grammar checking
Overall text layout			
Fluency and Structure	- Organisation in paragraphs	No. of paragraphs	Paragraph segmentation
Formal correctness	- Spelling errors	No. of spelling errors	Spell checking

Table 7: Example SFGL: summative assessment criteria

Analysis condition	Grade – Message
Communicative contents	
$TC = 6 \wedge LC = 4$	4 – Very good. You use the expected functions adequately.
(...)	
$TC \leq 3 \wedge LC \leq 4$	0 – Are you sure you have understood the purpose of this exercise?
Lexical contents	
$80\% \leq SV \leq 100\% \wedge 90\% \leq NW \leq 100\%$	4 – Excellent. Your text reads well and is precise. You are using the (...)
(...)	
$0\% \leq SV \leq 29\% \wedge 0\% \leq NW \leq 49\%$	0 – Careful. Your vocabulary is inappropriate and the text does not read well.
Sentence structure and accuracy	
$10 \leq NS \leq 9 \wedge 10 \leq NDM \leq 9 \wedge 0 \leq NGE \leq 1$	4 – Great. Your text is correct and adequate. There are no mistakes.
(...)	
$0 \leq NS \leq 7 \wedge NDM \leq 5 \wedge NGE \geq 3$	1 – Careful. Some information is missing and you are not using any connecting words.
Overall text layout	
$NP = 9 \wedge 0 \leq NSE \leq 1$	4 – Excellent. Your text has an adequate structure and no spelling mistakes.
(...)	
$NP \leq 5 \wedge NSE \geq 4$	0 – Careful: Your text does not have any structure and has many spelling errors.

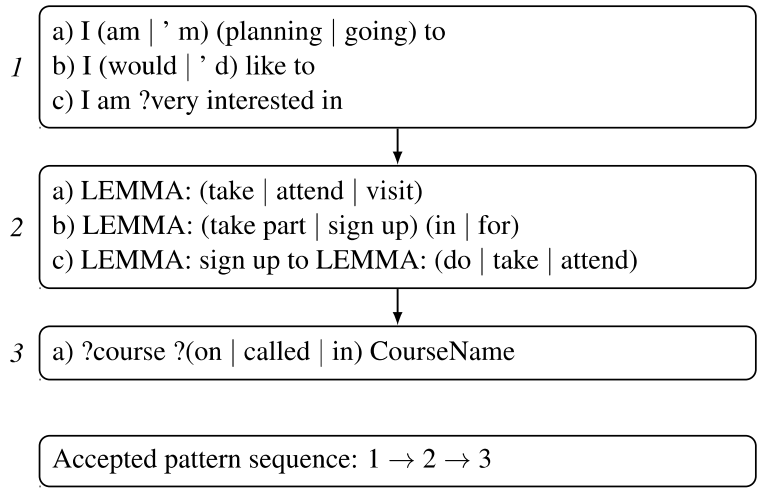

Figure 1: Patterns specifiable in finite-state formalism to recognize response fragment *Rel:CourseToTake*

Here, the question mark ('?') is used to indicate optionality and the vertical bar ('|') disjunction. In essence, the finite state formalism supports the very compact specification of a large (potentially infinite) set of strings. Using this method to characterize potential learner utterances supports the efficient identification of thematic and linguistic contents and to associate both well-formed and ill-formed subpatterns with feedback messages. The actual implementation of the NLP and feedback generation modules is discussed in detail in Badia *et al.* (2004) and Quixal *et al.* (2006). For an overview of the strategies most commonly used to analyze learner language and generate individualized feedback see Heift and Schulze (2007: Ch. 3), Meurers (2012), and Leacock *et al.* (2014).

Having exemplified our methodology to develop and design ICALL activities in a way that takes into account both pedagogical and computational requirements, we can now describe how the obtained task and NLP resource characterization helps validate the implementation of the task, as well as how the analysis of actual learner responses can inform the development of NLP corpus-based approaches.

4. Empirically validating task design

In the previous section we introduced a methodology for the development of ICALL tasks that integrates pedagogical and computational perspectives and needs. Maintaining this double perspective, this section presents a pilot

analysis of how the proposed fine-grained task and response characterization facilitates the evaluation of the class implementation of the task and the requirements of NLP resources. The goal is twofold: on the one hand, we want to illustrate that the methodology supports the empirical validation of the pedagogical goals by supporting the comparison of the theoretical predictions with the actual learner behavior. On the other hand, we want to argue for the use of such a methodology to inform the annotation of learner language, which is an essential part to design and develop appropriate and effective NLP-based assessment strategies. Generally speaking, the idea is to link the top-down task-based perspective of the proposed methodology to the bottom-up insights provided by a corpus-driven approach. We will first introduce the data that we use and then analyze authentic responses to the two tasks introduced in the previous section.

4.1. Learner data

The discussion is based on data collected at three different universities, with students who voluntarily took part in the process of piloting EFL learning materials enhanced with NLP-based automatic feed-back (Quixal, 2012: Chs. 7–9). For the pilot analysis included here, we focus on the two tasks introduced in the previous section. Seven students worked with the task *Stanley Broadband customer satisfaction questionnaire* at Heriot-Watt University (Edinburgh) in the spring of 2008, while 14 worked with the task *Registering for a course* at the Universidad Europea de Madrid and the Universitat Pompeu Fabra (Barcelona) in the spring of 2005. Participants were between 20 and 28 years old. Students at Heriot-Watt University were native speakers of six different languages (Arabic, Urdu, Galician/Spanish, Polish, Japanese and German), while students in Barcelona and Madrid were L1 speakers of Catalan and/or Spanish. According to the instructor, students in Edinburgh were at the B1 or B2 level. Students in Barcelona and Madrid qualified as B1 or beginner B2 level in DIALANG, a language level placement test (http://www.lancs.ac.uk/researchenterprise/dialang). All participants had been learning English for more than five years. They generally were experienced in using computers on a daily basis for studying, working, and searching for information.

Table 8 presents the basic information on the number and length of responses for both tasks. A total of 29 responses were collected for the customer satisfaction questionnaire task. The average response length for that task is 10 words and the standard deviation is close to 1 except for item 1, for which one of the responses was substantially longer. For the course registration task a total of 14 responses was collected, with an average length of 90 words and a standard deviation of 22 words.

Table 8: Basic information about the learner responses for the two tasks

Task	Total responses	Av. # of words	SD
Customer satisfaction questionnaire (All items)	29	10.7	2.4
Item 1	7	10	4.1
Item 2	6	13	1.0
Item 3	6	10	1.1
Item 4	5	10	1.1
Item 5	5	10	1.2
Course registration (E-Mail)	14	90.1	21.9

4.2. Evaluation of learner responses

Learner responses were automatically parsed with the NLP modules for the analysis of learner language sketched in Section 3.4. Next, the learner responses were manually reviewed to evaluate: (a) whether the contents of *Match* responses were correctly identified and whether the contents of *Alternative* responses could be classified into one of the specified response fragments; (b) the learner's accomplishment of the task's goals; and (c) the presence of ill-formed language. The first two steps were performed by one of the authors, the third was performed by an experienced EFL and English Linguistics professor at the Department of Translation and Language Sciences at the Universitat Pompeu Fabra. Going beyond the pilot study discussed for illustration purposes in this section, a thorough empirical study would include specific research questions, detailed annotation guidelines, and a multiple-annotator approach following standard procedures in the ICALL and NLP literature (Leacock *et al.* 2014; Meurers, 2015).

4.2.1. Overlap between responses and NLP specifications

Responses were classified as: (a) *Match*, if they were analyzed by the NLP module and thus considered envisaged by the AASF-based specifications; or (b) as *Alternative*, if they were not analyzed and thus considered as not envisaged.

4.2.2. Correctness of thematic content and well-formedness

Responses were classified as correct/incorrect given the task's thematic content, or as well-formed/ill-formed given their linguistic contents, resulting in four options: (a) correct thematic content and well-formed; (b) correct thematic content and ill-formed; (c) incorrect thematic content and well-formed; and (d) incorrect thematic content and ill-formed.

This kind of evaluation supports the core goal of an automatic feedback generation system, to implement feedback focusing on meaning and/or form. At the same time, the analysis supports the identification in the learner responses of the linguistic constructs targeted, e.g., to compute the linguistic complexity measures. In the following two subsections, as we present the pilot analysis, we will exemplify how the characterization obtained with the TAF and the RIF helps evaluate the learners' communicative and linguistic performance.

Exemplifying the categorization of learner responses. The response in (3) exemplifies a trivial *Match* (the specified pattern in *italics*) that is additionally classified as *Correct* and *Well-formed* response.

(3) What improvements would you like to see in the Stanley Broadband service?
What improvements would you like to see in the Stanley Broadband service?

The responses in (4) match with AASF-based linguistic patterns, despite being either incorrect and well-formed (IWF) or correct and ill-formed (CIF). Variation in (4a) is caused by a missing determiner and the noun *service* in plural form. Though the interrogative sentence is well-formed, the task's input introduces only one service whose name is *Stanley Broadband*. Variation in (4b) is a missing determiner resulting in an ill-formed sentence.

(4) a. What improvements you would like to see in IWF {Stanley Broadband services}?
What improvements would you like to see in ØDet S. B. LEMMA*: service?*
b. What improvements would you like to see in CIF {Stanley Broadband service}?
What improvements would you like to see in ØDet Stanley Broadband service?

The response in (5) is classified as Alternative, Correct and Well-formed (CWF). When compared to pattern (5i), variation results from the absence of the *addressee* (you), a different lexical choice (see vs. introduce), and the corresponding syntactic and semantic changes. When compared to pattern (5ii), it results from *improvements* being the subject of a passive sentence; and the shift in the thematic content that aims at increasing customer satisfaction rather than the people's will to subscribe.

(5) What improvements [CWF] {should be introduced to enhance customer satisfaction}?
   i) *What improvements would you like to see in the Stanley Broadband service?*
   ii) *What improvements would make more people want to subscribe to Stanley Broadband?*

4.2.3. Annotation of variation

Variation in learner responses was annotated too, both for well-formed and ill-formed variation. Annotations were classified into grammar (including spelling, morphology, syntax and semantics), sociolinguistics, text and functional knowledge. These coarse-grained categories correspond to Bachman and Palmer (1996)'s classification of linguistic knowledge and are sufficient to identify the nature of the variation in learner responses for the purposes of our research. In the following subsection, we exemplify these categories. A richer linguistics- and/or pedagogy-driven classification would be possible and, depending on the research question to be addressed, needed (Granger, 2003; Reznicek *et al.*, 2013; Meurers, 2015; Granger *et al.*, 2015).

4.3. Analysis of responses to the limited production task

4.3.1. Matching between envisaged and actual responses

Table 9 shows the number of foreseen (*Match*) and unforeseen (*Alternative*) responses for each item in the customer satisfaction questionnaire task. The first column shows the number of responses, the second one the number of patterns specified to analyze the learner responses. The third to the sixth columns show the patterns and the actual number of instances observed for each pattern depending on whether their class is *Match* or *Alternative*; alternative responses were grouped into patterns post hoc. Only for item 4 all learner responses matched one of the envisaged patterns. In contrast, none of the learner responses in item 3 used an envisaged pattern. Items 1, 2 and 5 elicited both envisaged and non-envisaged patterns.

4.3.2. Correctness of thematic content and well-formedness

Table 10 shows the distribution of correct and incorrect responses for each item. Comparing Table 9 with Table 10, we find that the ratio of match to alternative responses and the ratio of correct to incorrect responses are the same for items 2 and 4, but not for items 1 and 5. The proportion of correct/incorrect responses for item 3, for which no envisaged patterns were observed, is 2:4. While the numbers in this pilot study are too small to support general conclusions, the different ratios for different items can be taken to illustrate that learners may perform differently even within the same overall task, which has consequences both at the

Table 9: Observed foreseen and unforeseen patterns for the customer satisfaction questionnaire task

Item number	Total Responses	Specified Patterns	Match		Alternative	
			Patterns	*Instances*	*Patterns*	*Instances*
1	7	1	1	2	4	5
2	6	2	2	4	2	2
3	6	2	0	0	6	6
4	5	1	1	5	0	0
5	5	2	1	3	2	2
ALL	29	8	5	14	14	15

pedagogical and at the computational level. At the pedagogical level, such observations on a larger data set can inform the discussion on task difficulty, learner skills, or learner development. At the computational level, they highlight linguistic variation to be taken into account and increasing processing complexity.

Table 10: Correct and incorrect responses for the customer satisfaction questionnaire task

Item	Responses	Correct	Incorrect
1	7	3	4
2	6	4	2
3	6	2	4
4	5	5	0
5	5	2	3
ALL	29	16	13

Table 11 shows the breakdown of well- and ill-formed annotations across the other two classifications (Match/Alternative, Correctness of thematic content). We see that in general ill-formed variation is more frequent, independent of whether the responses were foreseen (*Match* vs. *Alternative*) or whether they include the expected thematic and linguistic contents (*Correct* vs. *Incorrect*). However, well-formed variation is higher for correct alternative responses.

Table 11: Observed well- and ill-formed structures for the customer satisfaction questionnaire task

Match				Alternative				Total	
Correct		*Incorrect*		*Correct*		*Incorrect*			
WF	IF	WF	IF	WF	IF	WF	IF	WF	IF
0	9	1	3	6	2	2	18	9	32

4.3.3. Observed variation

Quantitatively, nine of the 41 annotations are classified as well-formed variation, the remaining 32 as ill-formed. Well-formed variation occurs mainly in non-envisaged correct responses as in (5). On the basis of the RIF analysis we can interpret (5) as a correct response because as the instructions require: (a) it starts with *what improvements*; and (b) conveys the communicative goal, namely to learn from the customer how the service can be improved. This is variation at the functional level: A different exponent of function, with different lexical and syntactic choices that communicates felicitously in the context. The interplay between input data and communicative setting can be used to evaluate the learner's performance and the accomplishment of the pedagogical goals.

Out of the 32 instances of ill-formed variation, 27 of them belong to grammatical knowledge and five of them to sociolinguistic knowledge. Among the former, 13 are at the syntactic level, ten at the orthographic, three at the semantic, and three at the syntactic-semantic. Seen in Section 4.2.2, example (4a) illustrates ill-formed variation at the syntactic level (the determiner is omitted). Example (4b) illustrates ill-formed variation at the semantic level, since *services* and *service* would denote different world references.

Variation at the level of sociolinguistic knowledge is exemplified by (6) below, where the use of *thing* is deemed too informal and vague. The formal analysis of the setting as a company-customer communication in a professional context supports this type of evaluation and the annotation.

(6) What is the $^{\#}${thing} you like the least?

Note that the RIF analysis transparently supports the evaluation of the communicative goals as well as the pedagogical ones on the basis of linguistic evidence.

4.4. Analysis of responses to the extended production task

The 14 responses to the course registration task are emails, i.e., short texts. Given the lack of a modular item structure, we analyze them in terms of text fragments.

4.4.1. Matching between envisaged and actual responses

Table 12 shows the number of response fragments categorized as Match or Alternative. Two types of response fragments are shown: those corresponding to linguistic knowledge (LK) – formal aspects of writing an email (Greeting, Introduce yourself, Complimentary close and Signature), and those corresponding to thematic knowledge (TK).

Table 12: Observed foreseen and unforeseen patterns for the course registration task

Response Fragment	Specified Patterns	Match		Alternative		Missing
		Patterns	*Instances*	*Patterns*	*Instances*	
LK-Greeting	7	0	0	8	14	0
LK-IntroYourself	2	1	6	2	7	1
TK-YourDept	11	1	9	3	4	1
TK-Course	13	3	4	7	10	0
TK-Schedule	10	1	1	9	10	3
TK-AuthorisedBy	5	1	4	7	8	2
TK-UsefulFuture	6	2	4	3	4	6
TK-FutureInterest	6	2	2	5	7	5
LK-ComplClose	7	3	12	2	2	0
LK-Signature	1	1	12	0	0	2
ALL	61	15	54	38	66	20

The first column in Table 12 shows the number of patterns encoded in the rule-based grammars implemented on the basis of the task specifications, a sample response written by an EFL instructor and six responses written by learners of English for the purposes of writing the rules – see Section 3.4.

The second to the fifth column should be read in pairs: they respectively show the number of patterns foreseen (Match) or not foreseen (Alternative) and the number of instances observed for each pattern. Though the actual number of response fragments classified as Match and Alternative is not that different, note that: (a) there were many unobserved specified patterns – compare column 2 to column 3 – and (b) that the ratio of unforeseen patterns versus actual responses using them is often high – e.g., 8:14 for LK-Greeting and 9:10 for TK-Course. The last column shows that almost all response fragments were missing in at least one of the responses, and that thematic knowledge response fragments were more frequently omitted.

4.4.2. Correctness of thematic content and well-formedness

While Table 12 indicated a high number of fragments using non-envisaged patterns, Table 13 shows that there are a high number of correct response fragments. This is particularly the case for response fragments referring to thematic content, and less for those formally restricted by the text genre (except for *Greeting*). We here thus find that linguistic elements determined by the text genre or pragmatic contents, i.e., language knowledge, show less variation than those related to topical knowledge.

Table 13: Correct and incorrect responses to the course registration task

Item	Correct	Incorrect
LK-Greeting	1	13
LK-IntroYourself	13	0
TK-YourDept	12	1
TK-Course	13	1
TK-Schedule	10	1
TK-AuthorisedBy	11	1
TK-UsefulFuture	11	0
TK-FutureInterest	8	1
LK-ComplClose	8	0
LK-Signature	12	0
ALL	102	18

A qualitative analysis of the responses reveals well-formed variation at the level of exponents of function. (7) provides the specified patterns for the response fragment '*Course* to take' (the fourth row in Tables 12 and 13). Examples (8) and (10) show patterns actually observed in learner responses.

(7) a. I would like to sign up/register to/for take/do the course X.
b. I am interested in signing up/registering to/for take/do the course X.
c. I have signed up to take/do the course X.

(8) a. I am planning to take the X course.
b. I want to do the X course.

The patterns in (8) show variation with respect to (7a) at the lexical level in the main clause (*planning to* and *want to*), as well as at the subordinate level (omission of *register/sign up*). The function *asking someone to perform an action* can be realized by both patterns in (8), though the professional setting of the role-play activity could lead one to consider (8a) to be too direct or informal. Both responses present word order variation in the course denomination (*the X course* vs. *the course X*).

As an example of variation with respect to pattern (7a), the learner response in (9) shows the use of *writing to* describe the email's purpose. Though strictly speaking it is not requiring an action from the receiver, the expression is compatible with the task's setting and instructions and therefore it must be accepted as a correct pattern to express the *Course to take*.

(9) I am writing to you to register for the course on X.

The patterns in (10) exemplify variation with respect to (7b). (10a) shows again the omission of *register/sign up*, while (10b) and (10c), in addition to this omission, show a more complex linguistic structure including two juxtaposed simple sentences.

(10) a. I am interested in the course on X.
b. I am interested in one of your courses: namely the one on X.
c. I am interested in one of your courses. I am interested in the course on X.

4.4.3. Observed variation

Table 14 shows the distribution of ill-formed variation. Instances of ill-formed variation are lower in envisaged responses independent of their correctness: 22 (15+7) instances are found in 54 response fragments categorized as *Match*, and 81 (76+5) are found in the 66 fragments categorized as *Alternative*. It also shows that there is ill-formed variation in both correct and incorrect response fragments. In fact, it is more frequent in correct response fragments, 91 (15+76) vs. 12 (7+5).

Table 14: Ill-formed structures found in responses to the course registration task

Match		**Alternative**		**Total**
Correct	Incorrect	Correct	Incorrect	
15	7	76	5	103

Out of the 103 instances of ill-formed variation, 65% of them are related to grammatical contents, 20% to textual contents and 15% to sociolinguistic contents. Ill-formed variation at the level of grammar knowledge is found in errors related to syntactic issues, exemplified in (11): a preposition choice error (11a), and an agreement error (11b).

(11) a. I am NAME and I work **at* the Marketing Department.
b. I have no problem to take **this courses*.

There are also spelling errors such as the misspelling of *Thank* in (12a) and morphological errors such as the wrong formation of the past participle in (12b). We also find semantic errors as the use of *due to* to express cause instead of finality in (12c).

(12) a. **Than* you very much.
b. has **encourage* me to go on.
c. [It] could be interesting for my career *#due to* the marketing projects.

Finally, ill-formed variation at the level of sociolinguistics and of textual knowledge is exemplified by (13). The learner response (13a) is too informal for the setting, while in (13b) *and* is not appropriate in terms of cohesion.

(13) a. #*Hello*, (…)
b. I see my schedule and the timetable is fine with me #*and* I have the authorization of the department manager.

4.5. Discussion of the pilot analysis

Having characterized some of the aspects that arise from applying the analysis framework to the pilot data set, let us consider some conclusions from the empirical analysis we presented, both from a pedagogical and from a computational perspective. In doing so, we will argue that the detailed analysis of tasks that can be obtained by using our methodology substantially improves the opportunities to validate FL learning tasks and to assess their computational requirements and, in the end, their feasibility.

4.5.1. Pedagogical and language learning perspective

The explicit task characterization turned out to be useful for the interpretation, evaluation and annotation of learner responses. As illustrated by the discussion of the responses (4) and (5) for the task with a limited production response, the analysis readily supports a classification into well-formed or ill-formed independent of the classification into correct or incorrect. Similarly, we saw for the task with an extended production response that the patterns in examples (8) and (9) constituting variations at the lexico-syntactic level of the pattern in (7) were acceptable given the communicative setting of the target language use setting. Such analyses are directly relevant for the annotation of learner language and the formulation of target hypotheses (Lüdeling, 2008; Reznicek *et al.*, 2013; Meurers, 2015).

Turning from the general to the more specific insights, both of the analyzed tasks presented variation in terms of thematic knowledge and linguistic knowledge, although the one with a limited production response presented less variation at the level of thematic knowledge. This is intuitive because of two reasons: on the one hand, the customer satisfaction questionnaire task has a more direct and narrow relationship between input and response. On the other hand, the responses are shorter. In addition, the task with a longer response included much more input data to be processed, and, in fact, among the responses up to 17 response fragments (out of 20) were missing at the level of thematic content. This could speak for task complexity having an effect on variability of learner performance and the resulting product. The analysis suggests that lower task complexity and less margin for creativity produce less variation in learner responses, while increased task complexity and more

margin for creativity promote the use of self-chosen language and thematic knowledge.

In terms of linguistic knowledge, the data suggest that learners tend to make more errors when using functions and structures not foreseen by design. This could be explained in part because the foreseen patterns are structures that should be in principle primed by the input data and, of course, by the learning sequences in which the tasks are integrated. At the same time, independent of the type of response, variation was observed at different levels of linguistic description: grammar, sociolinguistics, text and functional knowledge. This may indicate that the tasks actually helped learners to engage in semi-structured language practice including linguistic and communicative skills.

The kind of analysis exemplified in the previous two paragraphs is relevant for FLTL and SLA. When based on a representative data set, the approach can also help empirically confirm whether learners are given the kind of tasks that leads them to practice the targeted learning goals. The analysis methodology can also be used to empirically substantiate and conceptually enrich classifications such as Littlewood's (2004). What kind of limited production response activities can be classified as communicative language practice? What kind of extended production response activities fosters the use of exponents for language functions that qualify as structured communication practice? Though our pilot study merely illustrates the approach and is too small to yield conclusive evidence, it confirms the applicability and usefulness of analyzing tasks and responses. Such analysis can support better informed decisions when it comes to task design and evaluation, a fundamental issue for CALL and technology-mediated Task-Based Instruction (Chapelle, 2001; Thomas & Reinders, 2010; González-Lloret & Ortega, 2014).

4.5.2. Computational perspective

Just as under the pedagogical perspective, the design-based specifications for the interpretation, evaluation and annotation of learner responses can be seen to be equally useful and relevant for the computational perspective. For example, the ratios of *Match* patterns and instances of patterns observed for the tasks are 5:14 (limited production) and 15:54 (extended production), while the ratios for non-envisaged patterns and their instances were 14:15 and 38:66. On the one hand, the ratios for envisaged patterns provide valuable information for the development of reliable NLP systems, where rule-based approaches can be successful, and where statistical or hybrid solutions are needed. In the project context we are building on, the analysis methodology successfully supported the design and implementation of ICALL activities by secondary school teachers using an authoring tool that automatically

generates the resources for the automatic assessment of learner responses by expanding an initial set of pre-envisaged responses (Quixal *et al.* 2010, 2012: Ch. 11). On the other hand, the results on non-envisaged patterns highlight the usefulness of corpus-driven approaches to developing individualized automatic feedback – an insight which is not yet established in the CALL community and that arguably would not only support better computational tools but also better design of pedagogical materials.

Two additional observations can be made here: First, the percentage of responses that actually used envisaged vs. non-envisaged patterns for the limited production task is around 50% each (14 vs. 15). In contrast, for the extended production response there were a total of 54 Match responses (12 of which are the signature of the email), 66 non-envisaged, 20 missing response fragments and six response fragments with unexpected contents – which were not analyzed but would have to be considered as a challenge for automatic processing since some of them might be additional but relevant, while others might be additional and irrelevant. These figures provide further support for the impact of response length on the complexity of learner language. They also highlight the empirical characterization of this complexity in linguistic and pedagogical terms as an important avenue for research.

Second, the analysis suggests that both well-formed and ill-formed variation occur frequently in learner responses, and variation was higher at the level of thematic knowledge. These findings support the need for research on robust content assessment for learner language, which can build on the growing interest in short answer assessment in NLP (Leacock & Chodorow, 2003; Ziai *et al.*, 2012).

While the data set we discussed is too small to support more detailed, statistically significant hypothesis testing, this pilot study exemplifies the approach with authentic data and confirms that activities that are pedagogically meaningful and computationally feasible can be successfully tackled on the basis of such a detailed analysis of tasks and learner responses. Applying such a systematic analysis to a range of language tasks completed by representative sets of learners can support a precise, empirical characterization of the viable processing ground for ICALL activities, a concept introduced in Bailey and Meurers (2008). Overall, the analysis thus supports an empirically informed strategy connecting (automatic) linguistic analysis and pedagogical goals in NLP-based CALL for task-based approaches.

5. Conclusions

To conclude, let us briefly sum up the main contributions of this paper. First, based on foundational research in SLA, FLTL and language testing, we proposed an enriched methodology for the design and implementation of tasks

and specifically ICALL tasks. We showed how this methodology provides a fine-grained characterization of learning tasks at the pedagogical and the computational level. Under such a characterization, learning activities can be classified based on the pedagogical purpose they serve and their computational feasibility. To the best of our knowledge, it is the first approach linking SLA, FLTL, and NLP needs based on a task-based approach to language instruction – one that supports both a conceptual top-down and an empirical bottom-up perspective on tasks.

Second, we presented an ICALL design framework that allows for the specification of NLP requirements given a set of pedagogical characteristics in a very detailed manner. The specification includes defining the criteria for correctness, expected linguistic structures, assessment criteria, and a feedback generation logic.

Third, we validated our methodology by applying it to two specific writing tasks with different pedagogical and linguistic characteristics. In spite of the small size of the pilot study, it illustrates that the information obtained from a fine-grained characterization of the learning tasks, a detailed specification of the linguistic and assessment needs, and a collection of learner responses can provide rich insights into the complexity and meaningfulness of ICALL tasks.

The variation in the learner responses we observed speaks for the combination of NLP approaches informed by design specifications and corpus-based analysis. While task specifications provide the patterns to handle a core set of learner responses, systematically integrating empirical generalizations and otherwise uncaptured exemplars in the observed learner responses to a given task can support a reliable broad-coverage automatic analysis.

Going beyond the use of the analysis framework for the generation of automatic feedback in ICALL, the perspective is equally applicable to the interpretation and annotation of elicited learner language in general. The methodology supports and resonates with the small but growing research strand aimed at the automatic analysis of linguistic structures in learner corpus data guided by the task's context (King & Dickinson, 2013). Supporting valid interpretation of learner data given explicit task and learner modeling (Meurers, 2015) as far as we see is equally relevant for the fields of SLA/FLTL and NLP and an opportunity for truly multidisciplinary research in CALL and ICALL.

Acknowledgements

The authors wish to thank Sowmya Vajjala and the anonymous reviewers for their valuable comments. The research presented was partially funded by: the ALLES project, 5th Framework Programme of the European Commission, contract number IST-2001-34246; a PhD student mobility grant of the Ministerio

de Educación del Gobierno de España, Subvención, TME2009-0266; and a paid research leave to Martí Quixal by Fundació Barcelona Media in 2010.

About the authors

Martí Quixal is a Research Associate at the Seminar für Sprachwissenschaft at the University of Tübingen and an Associated Researcher at the LEAD graduate school there.

Detmar Meurers is a Professor of Computational Linguistics at the University of Tübingen and a steering board member of the LEAD graduate school there.

References

Amaral, L., & Meurers, D. (2011). On using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. *ReCALL*, 23 (1), 4–24. Retrieved from http://purl.org/dm/papers/amaral-meurers-10.html. http://dx.doi.org/10.1017/S0958344010000261

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Badia, T., Díaz, L., Quixal, M., Ruggia, A., Garnier, S., & Schmidt, P. (2004). Individualised NLP-enhanced feedback for distance language learning. In Proceedings of ICALT. IEEE Computer Society, 2004. ISBN 0-7695-2181-9. URL http://www.computer.org/csdl/proceedings/icalt/2004/2181/00/21810729.pdf. http://dx.doi.org/10.1109/icalt.2004.1357638

Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In J. Tetreault, J. Burstein, & R. De Felice (Eds), *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, 107–115, Columbus, Ohio. URL http://aclweb.org/anthology/W08-0913. http://dx.doi.org/10.3115/1631836.1631849

Borin, L. (2002) What have you done for me lately? The fickle alignment of NLP and CALL. Paper presented at the EuroCALL 2002 pre-conference workshop on NLP in CALL, 14 August 2002, Jyvaskyla, Finland. URL http://k2xx.spraakdata.gu.se/personal/lars/pblctns/EuroCALL2002-NLP-WS.pdf

Brown, H. D. (2007). *Principles of language learning and teaching,* 5th edition. London: Pearson Education.

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139524681

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Estaire, S., & Zanón, J. (1994). *Planning classwork: A task-based approach. Educational Language Teaching*. Oxford: Macmillan-Heinemann.

González-Lloret, M., & Ortega, L. (2014). Towards technology-mediated TBLT. In M. González-Lloret, & L. Ortega (Eds), *Technology-mediated TBLT: Researching Technology and Tasks*, Volume 6, 1–22. Amsterdam/Philadelphia: John Benjamins.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal*, 20 (3), 465–480. URL http://purl.org/calico/Granger03.pdf

Granger, S., Gilquin, G., & Meunier, F. (Eds) (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139649414

Heift, T., & Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. New York: Routledge.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30 (4), 461–473. http://dx.doi.org/10.1093/applin/amp048

King, L., & Dickinson, M. (2013). Shallow semantic analysis of interactive learner sentences. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, Atlanta, GA USA. URL http://aclweb.org/anthology/W13-1702.pdf

Leacock, C., & Chodorow, M. (2003) . C-rater: Automated scoring of short-answer questions. Computers and the Humanities, 37, 389–405. http://dx.doi.org/10.1023/A:1025779619903

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.) vol. 25. San Rafael: Morgan & Claypool Publishers.

Littlewood, W. (2004). The task-based approach: Some questions and suggestions. *ELT Journal*, 58 (4), 319–326. http://dx.doi.org/10.1093/elt/58.4.319

Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In M. Walter & P. Grommes (Eds.), *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweispracherwerbsforschung*, 119–140. Tübingen: Max Niemeyer Verlag.

Meurers, D. (2012). Natural language processing and language learning. In C. A. Chapelle (Ed.), *Encyclopedia of Applied Linguistics*, 4193–4205. Oxford: Wiley, Oxford. URL http://purl.org/dm/papers/meurers-12.html. http://dx.doi.org/10.1002/9781405198431.wbeal0858

Meurers, D. (2015). Learner corpora and natural language processing. In S. Granger, G. Gilquin, & F. Meunier (Eds) *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press. http://purl.org/dm/papers/meurers-15.html. http://dx.doi.org/10.1017/CBO9781139649414.024

Nagata, N. (2009) Robo-Sensei's NLP-Based Error Detection and Feedback Generation. *CALICO Journal,* 26 (3), 562–579.

Nunan, D. (2004) *Task-based language teaching.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667336

Quixal. M. (2012). Language Learning Tasks and Automatic Analysis of Learner Lan-

guage. Connecting FLTL and NLP in the design of ICALL materials supporting effective use in real-life instruction. PhD thesis, Universitat Pompeu Fabra, Barcelona and Eberhard-Karls-Universität Tübingen. URL http://www.sfs.uni-tuebingen.de/~quixal/pubs/Quixal-12.pdf

Quixal, M., Badia, T., Boullosa, B., Díaz, L. & Ruggia, A. (2006). Strategies for the generation of individualised feedback in distance language learning. In *Proceedings of the Workshop on Language- Enabled Technology and Development and Evaluation of Robust Spoken Dialogue Systems of ECAI 2006,* Riva del Garda, Italy, September.

Quixal,M., Preuß, S., Boullosa, B. & García-Narbona (2010). AutoLearn's authoring tool: A piece of cake for teachers. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 19–27, Los Angeles, June. URL http://www.aclweb.org/anthology/W10-1003.pdf

Reznicek, M., Lüdeling, A., & Hirschmann, H. (2013). Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds), *Automatic treatment and analysis of learner corpus data*, Volume 59, 101–123. Amsterdam: John Benjamins. http://dx.doi.org/10.1075/scl.59.07rez

Richards, J. & Rodgers, T. (2001). *Approaches and methods in language teaching* (2nd ed.). Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511667305

Salaberry, M. R. (1996). A theoretical foundation for the development of pedagogical tasks in computer mediated communication. *CALICO Journal*, 14 (1), 5–34. Retrieved from https://webspace.utexas.edu/mrs2429/www/Salaberry1996CALICO.pdf

Schmidt, P., Garnier, S., Sharwood, M., Badia, T., Díaz, L., Quixal, M., Ruggia, A., Valderrábanos, A. S., Cruz, A. J., Torrejon, E., Rico, C. & Jimenez, J. (2004). ALLES: Integrating NLP in ICALL applications. In *LREC-2004. Conference on Language Re- sources and Evaluation*. http://www.lrec-conf.org/proceedings/lrec2004/pdf/3.pdf

Schulze, M. (2010). Taking ICALL to Task. In M. Thomas & H. Reinders (Eds), *Task-based language teaching and technology*, 63–82. London/New York: Continuum.

Thomas, M., & Reinders, H. (Eds) (2010). *Task-based language learning and teaching with technology*. London/New York: Continuum.

Weischedel, R. M., Voge, W. M., & M. James. An artificial intelligence approach to language instruction. *Artificial Intelligence*, 10 (3), 225–240. http://dx.doi.org/10.1016/S0004-3702(78)80015-0

Willis, J. (1996). *A framework for task-based learning*. Boston, MA: Longman Addison-Wesley.

Ziai, R., Ott, N., & Meurers, D. (2012). Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7) at NAACL-HLT*, 190–200, Montreal, June. Association for Computational Linguistics. http://aclweb.org/anthology/W12-2022.pdf