Towards a Corpus-based Approach to Modelling Language Production of Foreign Language Learners in Communicative Contexts

Voula Gotsoulia
Research Center for English Language
Faculty of English Studies
National and Kapodistrian University
of Athens
vgotsoulial@enl.uoa.gr

Bessie Dendrinos
Department of Language and Linguistics
Faculty of English Studies
National and Kapodistrian University
of Athens
vdendrin@enl.uoa.gr

Abstract

This paper discusses linguistic annotation issues, essential to a corpus-based approach to modelling the language use of foreign language learners in various contexts. We focus on learners of English and describe the corpora we use as well as the linguistic approach underlying their development. We present a scheme for describing grammatical choices and meaning components expressed in texts produced by learners. Our goal is to model the associations of corpus-attested linguistic patterns with their contexts, at different levels of language proficiency.

1 Introduction

Learning a foreign language is a complex process involving mastering a range of elements of a non-trivial system of communication and being able to use them appropriately in different social contexts. In a related vein, assessing a learner's ability to use language (i.e. his/her *linguistic competence*) is a significantly complicated task, requiring well-defined criteria for describing the instantiations of the system of language in socially meaningful ways. The *Common European Framework of Reference for Languages* (CEFR) has attempted to provide an objective basis for the explicit description of language proficiency across Europe, aimed to promote the transparency of language courses and 'the mutual recognition of qualifications gained in different contexts'.

CEFR distinguishes among several types of language-related communicative competences (i.e. lexical, grammatical, semantic, phonological, orthographic, orthoepic, sociolinguistic, pragmatic) and gives illustrative *descriptors* for each of these competences across the six-level scale of language proficiency established by the Council of Europe.[1] These descriptors are formulated in a very general way. In practice, incorporating their insight into concrete models of language learning and assessment is an open issue.[2]

In this paper, we address the foundations of a corpus-based approach to modelling the learners' production of language in relation to particular communicative contexts. Such a model can be used to support reliable assessment of language performance across proficiency levels, as well as test and materials development. We focus on English as a Foreign Language (EFL) and, more precisely, on the use of grammatical resources for the production of written texts.

In section 2, we describe the EFL learner corpora we use. Section 3 presents the linguistic framework we essentially draw upon and discusses methodological issues related to the representation of the range of grammatical resources employed by learners when producing written texts. Finally, in section 4 we specify the precise goals that we intend to pursue in the immediate future.

2 EFL Learner Corpora

As a basis for our study, we use the EFL learner corpora available from the KPG examinations, i.e. the Greek State examinations for certification of foreign language proficiency.[3] The KPG exams

[1] This scale comprises the European standard for grading language proficiency and includes the following reference levels: breakthrough or beginner (A1), waystage or elementary (A2), threshold or pre-intermediate (B1), vantage or intermediate (B2), effective operational proficiency or upper intermediate (C1), and mastery or advanced (C2).

[2] The English Profile Project, for instance, is currently working on providing concrete examples of the competences laid out in CEFR. It aims at clearly describing what a learner of English can be expected to know at each level (http://www.englishprofile.org/).

[3] The initials KPG stand for the Greek words 'Kratiko Pistopiitiko Glossomathias' (State Certificate for Language Proficiency): http://www.kpg.minedu.gov.gr/.

Proceedings of Recent Advances in Natural Language Processing, pages 557–561,
Hissar, Bulgaria, 12-14 September 2011.

(carried out since 2003) currently include six foreign languages (English, French, German, Italian, Spanish, and Turkish) and conform to the European scale of language proficiency.

Research carried out in the KPG project is related to the ongoing development of two databases, a database containing past papers and a database containing the candidates' answers and written texts (*scripts*). These databases are organised and linked to one another in terms of exam dates, languages, language levels, and exam modules. The scripts, in particular, are also classified in grading bands (i.e. fully satisfactory, moderately satisfactory, and unsatisfactory).

Our work will focus on written texts in the KPG script database for the English language. This corpus amounts to 3.5 million words and comprises collections of texts produced by learners of all ages in Module 2 (Written Production and Mediation) of the KPG exam. Module 2 tests a learner's ability to express himself/herself in written form by providing him/her with a *source text* as anchor to a particular communicative context and asking him/her to produce new texts in the target language (*target texts*). There are two types of source texts: one is in English and the other is in the candidate's mother tongue (Greek). In the latter case, the candidate is asked to *mediate* to an English speaker who does not speak Greek and relay the content of the source text, adapting it to a different context or a different communicative purpose.

The notion of *text* as the concrete configuration of discourse is central in the theory of language underlying the KPG exams. Departing from testing approaches emphasising the grammatical well-formedness of utterances, KPG emphasises the use of language as text in specific *contexts of situation* (i.e. communicative contexts). A *text* is defined as an independent unit of language which is meaningful for the context for which it has been produced. Put differently, it is a unit of language closely tied to aspects of a given situation (i.e. who is writing to whom, for what purpose, where the text might appear, etc.)

Both source and target texts stored in the KPG databases are described in terms of a number of parameters capturing information about their situational contexts (Kondyli and Lykou, 2010). These parameters include the text *type* (e.g. article, announcement, report, advertisement, prose excerpt, etc.), the *source* from which a text is taken (e.g. newspaper, magazine, encyclopedia, dictionary, web page, novel, etc.), the communicative *purpose* for which it has been produced (e.g. to inform, announce, convince, warn, invite, advise, protest, evaluate, etc.), the language *process* by means of which the purpose is fulfilled (i.e. description, narration, explanation, argument, instruction), the *domain* to which the text pertains (e.g. environment, travel, entertainment, science, sport, etc.), as well as the author's and addressee's communicative *roles* or identities (journalist, writer, friend, etc.).[4] Combinations of these parameters capture different text *genres*: a newspaper article written by a journalist who aims to inform readers about a scientific breakthrough by describing experiments, explaining goals, and arguing in favour of their importance differs from an article on the same topic published in a scientific journal, written by a scientist who aims to present his work to the academic community describing his experiments, explaining his goals and arguing in favour of the importance of his research.

Across the KPG exam levels, a variety of text genres and situations (ranging from everyday to formal communication) are associated with activities assessing different aspects of a learner's competence in the target language. The activities stored in the KPG databases along with the corresponding texts and their metadata and are managed and viewed via an intuitive web-based interface allowing SQL queries for information retrieval.

3 Corpus Annotation of Grammatical Patterns

The goal of our research is to describe in a systematic fashion the range of grammatical *choices* made by learners of English, using language in different communicative contexts, at different levels of proficiency. Furthermore, we seek to relate corpus-attested linguistic patterns with non-linguistic properties of the texts in which they appear, so as to model the contextualised use of language.

For this purpose, we also generalise across texts by organising the types of *text sources* currently

[4] *Processes* are defined in accordance with genre model proposed by Knapp and Watkins (2005). This model identifies genres that e.g. '*describe* through the process of ordering things into commonsense of technical frameworks of meaning, *explain* through the process of sequencing phenomena in temporal and/or causal relationships', etc.

specified by KPG in ontological structures. For instance, a novel, a short story, a fairy tale, a myth, a legend, a play script, and a comic strip are classified under a more general category called 'literary prose', which in turn inherits from a category referred to as 'literary text'; the latter is also inherited by 'literary rhythmic text' including poetry and lyrics. A newspaper, a magazine, and a news portal or blog are generally identified as 'news', while a letter, an e-mail, a note or comment, a postcard, and an invitation fall under the rubric 'interpersonal communication text'. In a similar way, text types, domains, and types of authors and addresses are also organised ontologically. This sort of classification can support the study of language use across generalised situational contexts.

The linguistic framework which we adopt for modelling language use is Halliday's Systemic Functional Grammar (SFG) (Halliday, 1976, 1985). Functional linguistics emphasises the continuities between language and social experience (i.e. real-world situations). That is, SFG views language is a system of *semiosis* that cannot be divorced from its context. It describes the resources of this complex system in terms of a compositional structure comprising three distinct layers (*strata*): *phonology*, *lexicogrammar* and *semantics*. Lexicogrammatical resources create meaning in the form of *text*.

3.1 The Annotation Scheme

Annotation of grammatical patterns spans across four types of text units: *sentences*, *clauses*, *phrases*, and *words*. For each text unit, we distinguish two levels of linguistic description: a Grammatical Type (GT) and a Semantic Type (ST) level. The former includes morphological and syntactic information about the unit in question, while the latter describes its semantic *function*, i.e. its function as a building block of textual meaning.

To illustrate the scheme with a concrete example, consider the sentence (1), taken from a B2 level script.

(1) I read in your email that you are thinking to quit school and work as a waitress, because you want to make money and travel all over the world.

The annotation of (1) involves several *annotation sets*. Each one includes combinations of GT and ST labels for a given type of text unit. The set shown in (2) describes the whole sentence.

(2) GT: S.Complex
ST: S.Declarative

Following the insight of Systemic Functional Grammar, we classify sentences in one of the types: Simple, Compound, and Complex. A sentence is an independent utterance with complete meaning. A Simple sentence typically contains a verb and its arguments.[5] A Compound sentence comprises two or more interdependent clauses of equal status (e.g. *[He came to a thicket] and [at that time he heard the faint rustling of leaves]*) (the definition of the clause follows). A Complex sentence includes two or more interdependent clauses of unequal status (e.g. *[When the path reaches the road], [follow the road downhill for about 200 metres]*).[6]

At the semantic level, we classify sentences as *Declarative*, *Interrogative*, *Imperative*, or *Exclamatory*. These categories essentially capture what Halliday (1979) called the *interpersonal* function of language referring to the ways in which meaning is negotiated between participants in a communicative act.

Another annotation set for (1) includes the descriptions in (3), (4), and (5) below, representing the clauses '*I read in your email*', '*that you're thinking to quit school and work as a waitress*', '*because you want to make money and travel all over the world*', respectively. A clause is a dependent utterance with incomplete meaning; it comprises a verb and its subject (at least).

(3) GT: Cfin_act.Main
ST: C.Mental

(4) GT: Cfin_act[that].Dep_Obj
ST: C.Mental

(5) GT: Cfin_act[because].Dep
ST: C.Mental

These representations capture the grammatical properties of the clauses above as well as their semantic functions. The utterance in (1) involves three clauses with finite (*fin*), active voice (*act*) verbs. (3) is the main clause, which introduces the semantic basis of the utterance. The semantics of (4) depends on that of (3) (i.e. it is the Object

[5] Yet an utterance like 'Hello!' or, simply, an exclamation is also considered a Simple sentence.

[6] The examples in parentheses are from Halliday and Matthiessen (2004). The different degrees of interdependency between sentences are referred to with the terms *parataxis* (equal status) and *hypotaxis* (unequal status).

of its verb), while (5) depends on (4). The structural (syntactic) typing of clauses (i.e. Main, Dep, Dep_Subj, Dep_Obj) is recorded at the GT level.

The semantic level comprises a description of the content of each clause. The content is represented in terms of general types of *events* or *processes*, as identified by Halliday (2004), i.e. *Mental*, *Verbal*, *Material*, *Relational*, *Behavioural*, *Existential* processes. We define these processes as functions referring to real-world events or situations. For their definitions, we specify sets of properties shared by participants in the designated events or situations. Note that we replace the Material type (whose definition is somewhat vague) with a Causation type (referring to events with causally affected participants) and we include additional types: Intentional Action, Motion, and Possession (see Gotsoulia (2011) for a description of the theoretical approach we adopt for defining broad categories of event semantics).

Similar annotation sets are specified for Verb Phrases (VPs), which also denote events, as exemplified by the representations of the phrases '*to quit school*' (6), and '*travel all over the world*' (7):

(6) GT: VPinf_act[to].Dep_Obj
ST: VP.Intentional_action

(7) GT: VPinf_act[to].Dep_Obj
ST: VP.Intentional_action

As illustrated in the above representations, our scheme emphasises the significance of general events in the creation of textual meaning. The linguistic expression of events is captured across different types of text units (i.e. clauses and phrases). Note that at the phrase level, we also represent Noun Phrases (NPs) (i.e. nominalisations), Adjectival Phrases (ADJPs), or Prepositional Phrases (PPs) denoting events of the sort we are interested in:

(8) [$_{NP}$The announcement of the results] was postponed. (*GT:NP, ST:NP.Verbal*)

(9) He is [$_{ADJP}$interested] [$_{PP}$in working] as a translator. (*GT:ADJP, ST:ADJP.Mental*) (*GT:PPing, ST:PP.Intentional Action*)

4 Future Work

The two-layer annotation scheme presented above encodes systematic associations of criterial lexical functions forming *textual meaning* and grammatical structures expressing each function.

Currently, we are in the process of annotating a portion of the KPG corpora with SFG categories. From the annotated data, we will be able to acquire frequencies of lexicogrammatical patterns in particular communicative contexts, proficiency levels, and grading bands. The novelty of our approach lies exactly at the combined representation of lexical and grammatical components, which (to our knowledge) has not yet been explored in the analysis of learner corpora. For example, the relevant research strands in the English Profile Project (i.e. the morpho-syntactic and the lexico-semantic strand) are unrelated.

While annotation is currently carried out manually, in the immediate future we intend to address semi-automatic tagging of SFG lexicogrammatical categories by using a syntactic and a semantic parser and mapping the output to the designated SFG categories. The proposed representations can ultimately be used to support reliable, semi-automatic assessment of contextualised language use in learners' scripts by computing similarities of graded and novel (not graded) scripts in terms of lexicogrammatical features and their frequencies.

Acknowledgments

Work presented in this paper is part of the project 'Differentiated and graded National Foreign Language Examinations', co-financed by the European Social Fund and National Resources.

References

Buttery, P. 2009. *Using large-scale corpora within the English Profile program; computational methods for constructing reference levels descriptors.* AAAL 2009. Denver, Colorado.

Gotsoulia, V. 2011. *An abstract scheme for representing semantic roles and modelling the syntax-semantics interface.* In Proceedings of the 9th International Conference on Computational Semantics (organised by the ACL Special Interest Group on Computational Semantics). Oxford, United Kingdom.

Halliday, .A.K. and Hasan R. 1976. *Cohesion in English.* London: Longman.

Halliday, M.A.K. and Hasan R. 1985. *Language, Context and Text: Aspects of Language in a Social-semiotic Perspective.* Oxford: Oxford University Press.

Halliday, M.A.K. and Matthiessen C. 1999. *Construing experience through meaning.* London, New York: Continuum.

Halliday, .A.K. and Matthiessen C. 2004. *An Introduction to Functional Grammar. (3rd ed.)* NY: Arnold.

Hawkins, J.A. and Buttery, P. 2009. *Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme.* In: Studies in Language Testing: The Social and Educational Impact of Language Assessment. Cambridge University Press.

Knapp, P. and Watkins, M. 2005. *Genre, text, grammar: Technologies for teaching and assessing writing.* Sydney: UNSW press.

Kondyli M. and Lykou C. 2010 *Linguistic Description of the KPG tasks and texts: The text type and lexicogrammar perspective. (in Greek)* Research Periodical: http://rcel.enl.uoa.gr/periodical/article$_e$n.htm