EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 191–201
CILC2016. 8th International Conference on Corpus Linguistics

SEPAME2: The first longitudinal corpus for Greek as an L2

Maria Iakovou, Olga Dima, Tatiana Katsina, Maria Kavvadia, Sophia-Nepheli Kitrou, Christina Kostakou, Marina Koutsoubou, Froso Pappa, Stavrialena Perrea, Irianna Vasileiadi-Linardaki & Flora Vlachou

National and Kapodistrian University of Athens
sepame2@phil.uoa.gr, sepame2@gmail.com

Abstract

SEPAME2 is the first attempt to design and implement a longitudinal corpus of different L1 learners of Greek as an L2. It supports the idea that the best way to learn a language is by being "pushed" to use it in different circumstances/registers and by taking advantage of personalized feedback modes, so that the language becomes not only the result of the learning process, but also the source of further metalinguistic reflection. In this preliminary presentation, main design principles as well as future implications of the SEPAME2 project are discussed.

1 Introduction

Learner corpora, defined as electronic collections of written or spoken texts produced by language learners (Granger 2008: 259), can contribute to Second Language Acquisition (SLA) theory and research by enlightening the cognitive processes of language learning and providing a more accurate description of interlanguage. Current learner corpora tend to be synchronic, which means they describe learner use at a particular point of time. There are very few longitudinal corpora, that is, corpora which cover the evolution of learner use. For Granger (2002:11) "the reason is simple: such corpora are very difficult to compile as they require a learner population to be followed for months or, preferably, years". The challenge becomes even greater, when the research objective focuses on one of the less widely spoken and taught languages, such as Modern Greek, where there are no longitudinal corpora at all. The learner corpus research is eliminated in some cross-sectional corpora quite easily compiled by language certification texts or written essays produced for special occasions, such as those produced by young learners participating in educational programs, like Diapolis or Muslim Minority Education (Tzimokas 2010, Kiliari 2014, Tantos 2015).

SEPAME2, which stands for the Greek initials of *Learner Longitudinal Corpus for Greek as a Second Language (L2)*, was initiated in October 2014 at the Modern Greek Language School of the

National and Kapodistrian University of Athens. It is the first attempt for Greek as an L2 to cross the bridge between corpus linguistics theory and second language practice. The aim of the project is to build a large longitudinal corpus of written and oral texts derived by different L1 background adult learners, and, therefore, contribute to filling the gap in corpus-based SLA studies. In the SEPAME2 project the same students are followed over a period of at least one academic year and data collection is organized once a month. All learners' productions receive linguistic and metalinguistic comments as part of a personalized feedback service provided by our team. In this sense, feedback creates both the necessary incentive for all SEPAME2 participants and sufficient assistance to "push" their output further. In fact, this is the exact translation of SEPAME in Greek (= pushing learner's abilities further). These three features, longitudinal dimension, duality of mode in language production and personalized feedback service, provide the innovative framework on which our proposal is built.

2 Methodology

In this section we present the number of participants and the criteria they have to fulfill in order to become members of our research (2.1), the data collection process (2.2), and the nature of the task design (2.3).

2.1 Participants

SEPAME2 is envisaged as a developmental learner corpus of written and spoken texts, stored in an electronic format, to be used for interlanguage analysis and other pedagogical applications. So far, it comprises 439 learners who have been monthly followed during the academic years 2014-2015 and 2015-2016 (Table 1). It is important to notice that 40 out of 345 learners who participate during the current academic year in our project have been followed since the initial SEPAME2 stage.

Level	**Initial SEPAME2 stage**	**Current SEPAME2 stage**	
	2014 - 2015	**2015 - 2016**	**Total**
A1-A2	41	180	**221**
B1-B2	35	103	**138**
C1-C2	18	62	**80**
Total	94	345	**439**

Table 1: Number of participants

All participants have to fulfill the following criteria:

- They have to attend annual language courses at the Modern Greek Language School of the National and Kapodistrian University of Athens (that is, October to May, for 3 hours daily).

- They have to be followed over a period of at least one academic year.
- They have to pass the School placement test in order to be placed at one of the six language levels (A1-C2), according to the Common European Framework for Languages (2001).

2.2 Data collection process

The learners consent to participate voluntarily in the project by signing a consent form with all relevant information about the study and the data confidentiality (Mackey & Gass, 2005: 25-36). They also fill in a questionnaire with detailed information about their profile as learners. In particular, the details provided refer to:

- Personal information: age, sex, nationality, mother tongue, Greek relatives, if applicable, and languages spoken at home.
- Exposure to Modern Greek: length of stay in Greece and information regarding Greek language courses attended at the past, if applicable -name of the institution (school or university) where classes were held and courses' length.
- Educational and language background: level of educational attainment, other languages spoken and self-rating of proficiency in each of these languages.
- Motivation related to the Greek language learning.

These different variables, especially learners' L1 and their motivation to learn Greek, are essential for fine-grained, quantitative analyses, and can shed some light on unexpectedly strange results, counter-intuitive and conflicting with established descriptions (Sinclair 2005, Lozano & Mendikoetxea 2013). The SEPAME2 corpus contains data from learners of 36 different mother tongue backgrounds.

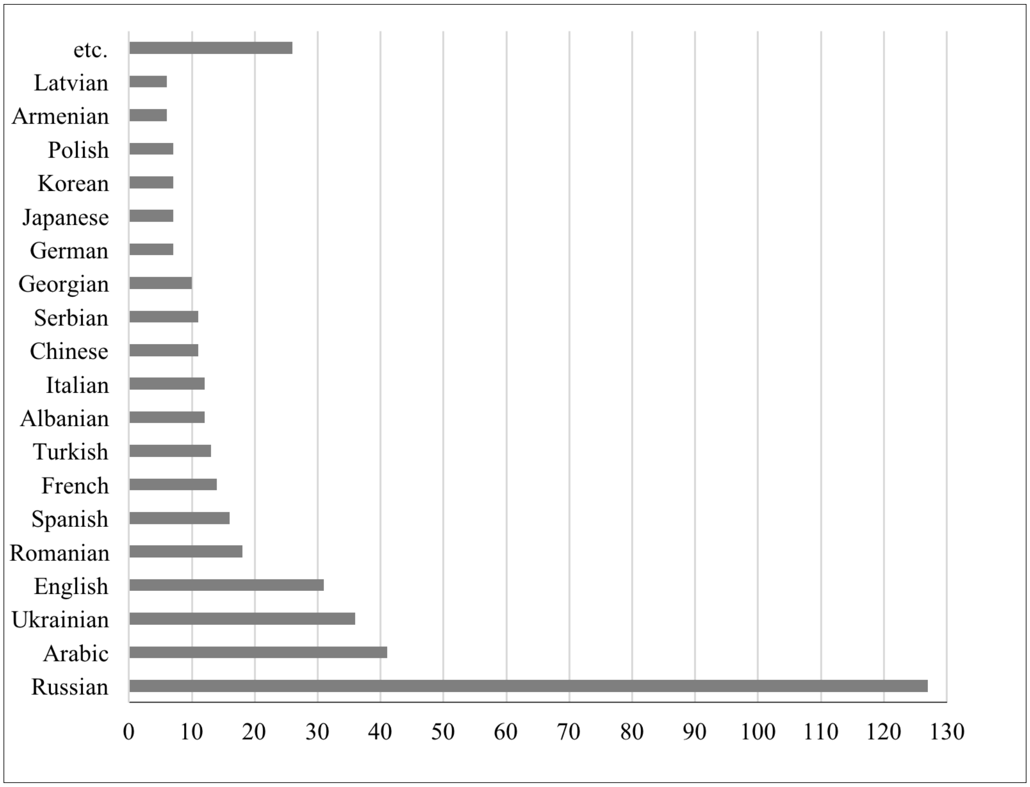

Figure 1: Participants' L1

Figure 1 shows that the great majority of participants speak Russian. The languages that are not presented above are those that are spoken by less than 4 native speakers, such as Icelandic, Slovenian, Swedish, etc.

As far as the learners' motivations are concerned, they are divided in two categories, according to Saville-Troike's division (2006: 135). BAS.I.C. describes the basic interpersonal competence, such as family, tourism, and personal interest, and COG.A.C. refers to the cognitive academic competence, such as employment and studies.

Regardless of the language level, the levels of both categories are very high (Figure 2). However, in level A COG.A.C. is higher but this is due to the fact that one of the major goals of the Modern Greek Language School is to prepare the learners for their academic studies in the Greek Universities.

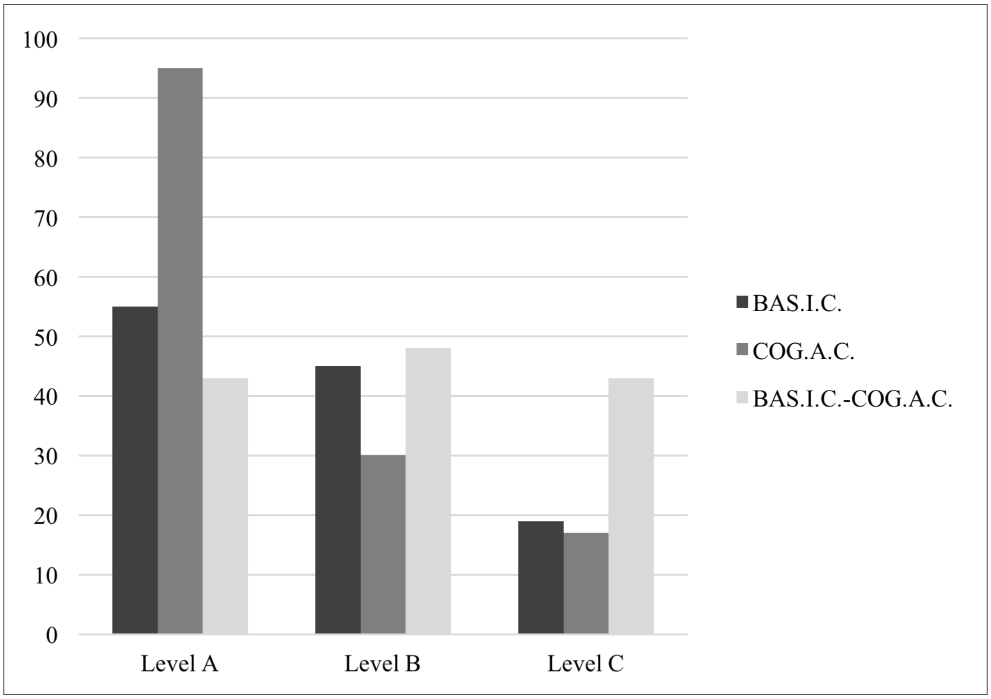

Figure 2: Participants' motivations

The implementation of our project involves a series of developmental data drawn from different sources and distributed in one of the following «pools»:

- **Pool 1** comprises handwritten texts produced during the regular school schedule in class or at home, but under the teachers' supervision. All texts are digitized and chronologically stored with an electronic format comprising metadata information, such as the writing conditions, the target genre (+/- formal style, +/- recipient), and the use of reference materials (dictionaries, textbooks, etc.).
- **Pool 2** includes handwritten texts produced during the extra writing practice courses organized by our team and offered to all project participants. In the predetermined time of 30 minutes, each learner has to write monthly (extra writing practice courses open usually on the last week of the month) a task related to his/her language level. Therefore, the expected task production in the period of one academic year (October to May) is about 7 tasks per learner. All productions are digitized and e-mailed back to the learners with a personalized linguistic and metalinguistic feedback for all kinds of errors (grammatical, lexical, and pragmatic). Moreover, all texts are edited and presented anew in an error free form, so that each learner may benefit from the second draft of his/her initial production (Ellis & Barkhuizen, 2005).
- **Pool 3** comprises oral productions by the same learners on similar tasks as those described in the previous pool. Oral data elicitation takes the form of informal interaction between learners and native speakers (our research team's members), and occurs twice a year for 10 minutes (1st

collection: December to February, 2nd collection: April to May). Transcription of spoken material is broad orthographic, marking basic features of spontaneous discourse such as overlap, pauses, interruption, lengthening, etc. and it allows us to compare both oral and written output of the same learner. A digital copy of all spoken texts allows for more detailed transcriptions when the need arises in the future. Furthermore, a direct feedback form assessing the learners' discourse competence as well as their grammatical and lexical inefficiencies is provided by the end of any spoken interaction.

- **Pool 4** refers to the final storage of the learners' performance and it contains written and oral material drawn from the School achievement test, known as *Certificate of Greek Language Knowledge* which corresponds to the B2 level (CEF, 2001) and is held every May.

Therefore, each subject participating in the project is followed under different output conditions (+/- testing pressure, +/- teacher intervention) and may inform the corpus with two final texts, one oral and one written, assessed by external evaluators.

In this article we focus on the quantitative data drawn from Pool B and C, which consist the core of the SEPAME2 corpus. They consist of data generated by our own tasks (presented below) and collected regularly in a monthly or so basis. Moreover, it is on this kind of output that different feedback types (+/-direct, linguistic, metalinguistic) are applied and the results of this intervention will be evaluated in relation to the learners level and their developmental progress.

		2014-2015			**2015-2016**			**Total**
		A1-A2	**B1-B2**	**C1-C2**	**A1-A2**	**B1-B2**	**C1-C2**	
pool 2	number of texts	246	216	90	614	268	192	**1626**
	number of words	25.819	32.147	14.160	37.142	29.510	21.352	**160.130**
pool 3	number of interviews	74	70	20	140	72	50	**426**
	number of words	97.714	83.749	12.500	184.864	86.142	31.250	**496.219**

Table 2: Number of data

As mentioned in section 1, SEPAME2 data collection started in 2014. Data are still being collected, so that our corpus will eventually contain more than one million words. As shown in Table 2, 1.626 written texts and 426 interviews have been collected to date (February 2016). In addition, our corpus already consists of more than 650.000 words. Our corpus size and the strict design criteria on the basis of which it has been collected (cf. sections 2.2 & 2.3) consist major assets in terms of representativeness of the data and generalizability of the results[1], setting the ground for new reference tools and more reliable language data.

[1] For key design criteria in learner corpus compilation, see Granger (2008) and Sinclair (2005).

2.3 Task Design

Learners' productions are the outcome of graded tasks concerning entities such as people, places, events, and situations. Moreover, these tasks encompass a variety of topics (everyday life, health, environment, etc.), different text types (personal and formal letters, informal interviews, free compositions, etc.) and different genres (description, narration and argumentation) that learners have to perform in relation to their language level and their cognitive abilities (A1-C2, according to CEF, 2001). Therefore, our tasks can potentially elicit all possible linguistic structures and a wide range of vocabulary aiming to a high degree of inclusiveness and representativeness. The basic pedagogic claim is that language tasks should be sequenced on the basis of the concepts that the task requires, in order to be expressed and understood (e.g. relative time, spatial location, causal relationships, and intentionality). This claim is mainly advocated by the Cognition Hypothesis (CH, Robinson 2001, 2011, Robinson & Gilabert 2007), which predicts that increasing task complexity influences the quality of second language production and creates the conditions for further L2 learning and interlanguage development. For example, tasks requiring simple description of events happening now (+ here-and-now), in a shared context (+ familiarity), where few elements (+ few elements) have to be described are less cognitively and so less linguistically demanding than tasks requiring reference to events that happened elsewhere, in the past (- here-and-now), where many elements have to be distinguished (- few elements), and where reasons have to be given to support statements made (+ reasoning). Moreover, it is important to note that our task design and implementation apply for the very first time cognitive features on the language levels description. Therefore, all tasks match the CEF guidelines with the CH conditions, as it can be illustrated by the example below (Table 3). Table 4 shows successive A level

Pool B: Writing			
Level:	**A1-A2**	**B1-B2**	**C1-C2**
task 1 [October]	**Person description** related to the learners' self Guided production by answering formulaic questions (name, age, family situation, languages, job) + here-and-now - reasoning + few elements	**Person description** out of the learner's self (physical appearance, likes) Semi-controlled production by answering open-ended questions + here-and-now - reasoning +/- few elements	**Person description** in and out of the learner's self (personal identity, likes and dislikes leading to choices) No-controlled production based on the learner's reasoning + here-and-now + reasoning - few elements

Table 3: Complexity grading among the language levels

tasks and reflects how the complexity is increased not only among the different language levels (A-B-C), but among the seven tasks of the same language level, as well.

A1-A2	task 1 [October]	Person description related to the learners' self	Guided production by answering formulaic questions (name, age, family situation, languages, job)	+ here-and-now - reasoning + few elements
	task 2 [November]	Person description out of the learner's self (personal identity, physical appearance, likes)	Guided production by answering formulaic questions (name, age, family situation, likes, dislikes)	+ here-and-now - reasoning + few elements
	task 3 [December]	Plan presentation related to the learner's self	Guided production by answering open-ended questions (habits, likes, dislikes)	+ here-and-now - reasoning + few elements
	task 4 [January]	Event narration related to the learner's self	Guided production by answering open-ended questions (with spatial, temporal, relational reference)	- here-and-now - reasoning +/- few elements
	task 5 [February]	Place description out of the learner's self + comparison + personal assessment	Guided production by answering open-ended questions (concerning the spatial comparison of the concrete pictures' properties)	+/- here-and-now + reasoning + few elements
	task 6 [March]	Problem-solving out of the learner's self	Guided production focused on advice giving	+/- here-and-now + reasoning + few elements
	task 7 [April]	Hypothetical narration related to the learner's self	Guided production focused on the learner's personal desires, dreams, emotions	- here-and-now - reasoning + few elements

Table 4: Tasks' complexity grading sample (Level A)

With regard to the genres of the tasks as a whole, in A1-A2 levels there are four descriptive tasks, whereas in C1-C2 four out of seven tasks require learners to perform argumentative writing topics (Table 5).

	A1-A2	**B1-B2**	**C1-C2**
Description	4	3	1
Narration	2	3	2
Argument	1	1	4

Table 5: Tasks' genres

	A1-A2	**B1-B2**	**C1-C2**
- here-and-now	**2**/7 tasks	**3**/7 tasks	**4**/7 tasks
- few elements	-	**3**/7 tasks	**7**/7 tasks
+ reasoning	**2**/7 tasks	**4**/7 tasks	**6**/7 tasks

Table 6: Cognitive features

As shown in Table 6, the complexity is increased in C1-C2 levels, where most of the tasks require reference to events happening elsewhere, in the past or the hypothetical future, (**-** here and now), where many elements have to be distinguished (- few elements), and where reasons have to be given to support statements made (+ reasoning).

On the contrary, concerning A levels, only two out of seven tasks require reference to minus here-and-now and reasoning features, as well as there is not a single task demanding many elements to be combined. This arises from the necessity of less cognitively and so less linguistically demanding tasks, i.e. simple description of events happening now, where few elements have to be described.

3 Implications, applications and prospects

Halliday (1993: 4) asserts that corpus linguistics "re-unites data gathering and theorizing and this is leading to a qualitative change of our understanding of language". This view is echoed by our SEPAME2 project, whose the major implication seems to open new ways in learning and teaching of

Greek as an L2. To this end, future plans include first, the completion of data collection and the implementation of this new reference corpus with free access to all teachers and learners, who may benefit from successively produced texts, and second, the dissemination of the personalized feedback service provided, so that all types of errors will be analyzed and commented. Moreover, one of the main priorities of SEPAME2 includes the tagging of the corpus, which is not yet fully tagged, besides some samples that have been preliminary annotated. For this reason, the next academic year's research plan involves the adaptation of tagging tools, such as *ELAN* (Varlokosta et al. 2016), *ILSP _NLP (WSDL), Episimiotis* (Tzimokas 2004), to the annotation of our data.

Finally, according to our view, this approach on the learner corpus design may be beneficial to:

- The researchers' community as a whole: Our data will allow linguists to conduct interesting and reliable corpus-driven studies for the interlanguage system of learners of Greek.
- The applied linguists: Our data will be helpful for Greek as L2 teachers, syllabus designers, task designers, language test providers and textbook writers in order to draw useful conclusions about the learning stages of different L2 learners and to raise their awareness about the realizations of different language levels.
- The learners of Greek as an L2: They will take great advantages in order to improve their productive skills' performance as long as they study Greek as an L2 at the National and Kapodistrian University of Athens and are exposed to the systematic feedback mode offered by our team.
- The post-graduate students of the Interdepartmental Program of Greek as an L2 organized by the National and Kapodistrian University of Athens: The project will familiarize them with methods and data highly frequent in the fields of Second Language Acquisition and Foreign Language Teaching.
- The computer experts: The developmental learners' data are a prerequisite source of data for the construction of new language tools, such as L2 automatic spellers, L2 taggers, error encoders etc.

SEPAME2 aims to provide evidence for a more comprehensive, accurate and authoritative description of the Greek learners' language. Our team is devoted to this aim, though we are aware that this is just the beginning!

References

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching and assessment*. Cambridge: Cambridge University Press.

Ellis, R. & G. Barkhuizen (2005). *Analyzing Learner Language*. Oxford: Oxford University Press.

Grabe, W., & Stoller, F. L. (1997). "Content-based instruction: Research foundations". In M. A. Snow, & D. M. Brinton (eds.) *The content-based classroom: Perspectives on integrating language and content* (5-21). NY: Longman.

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (eds.), *Corpus linguistics. An international handbook*, *1*, (259–275). Berlin and New York: Mouton de Gruyter.

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (eds.), *Computer learner corpora, Second Language Acquisition and Foreign Language Teaching* (3-33). Amsterdam: Benjamins.

Halliday, M.A.K. (1993). Towards a language-based theory of language. *Linguistics and Education*, 5: 93-116.

Kiliari, A., Archakis, A., Tsakona, V. (2014). Learner corpora. In *Narratives: Foreign learner corpora*. Research project "Narratives, tracing identities and linguistic intervention" of the Aristotle University of Thessaloniki. [In Greek]

http://www.del.auth.gr/index.php/el/erevna/ergasthrio-ekpaideftikhserevnas/afigiseis.

Lozano, C. & Mendikoetxea, A. (2013). Learner corpora and Second Language Acquisition: The design and collection of CEDEL2. In N. Ballier, A. Díaz-Negrillo, & P. Thompson (ed.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins.

Mackey, A. & S. M. Gass (2005). *Second Language Research: Methodology and Design*. Mahwah/ New Jersey/ London: LEA.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (ed.), *Cognition and Second Language Instruction* (287–318). Cambridge: Cambridge University Press.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (ed.), *Second Language Task Complexity: Researching the Cognition Hypothesis of language learning and performance* (3-37). Amsterdam/Philadelphia: John Benjamins.

Robinson, P. & R. Gilabert. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching*, 45 (3), 161–176.

Saville-Troike, M. (2006). *Introducing Second Language Acquisition*. Cambridge: Cambridge University Press.

Sinclair, J. (2005). How to build a corpus. In M. Wynne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (79-83). Oxford: Oxbow books.

Stoller, F. L. (2002). Content-Based Instruction: A Shell for Language Teaching or a Framework for Strategic Language and Content Learning? Keynote presented at the annual meeting of Teachers of English to Speakers of Other Languages, Salt Lake City. (Online at CoBaLTT website)

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass and C. Madden (eds.), *Input in second language acquisition* (235-253). Rowley, MA: Newbury House.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J.P. Lantolf (ed.), *Sociocultural Theory and Second Language Learning* (97-114). Oxford: Oxford University Press.

Swain, M. (2001). Integrating language and content teaching through collaborative tasks. *Canadian Modern Language Review*, 58 (1), 44-63.

Tantos, A. (2015). «Learner corpora in second/foreign language learning». In D. Papadopoulou, E., Agathopoulou, & K. Pouliou, (eds.), *Υποστήριξη της Λειτουργίας των Τάξεων Υποδοχής: Ζητήματα γλωσσικής διδασκαλίας* (297-329). Aristotle University of Thessaloniki. [In Greek]

Tzimokas, D. (2010). Learner corpus for Greek as a Second Language: A new research and instruction tool. *Studies in Greek Linguistics*, 30, 602-616. [In Greek]

Varlokosta, S., Stamouli, S., Karasimos, A., Markopoulos, G., Kakavoulia, M., Nerantzini, M., Pantoula, Ai., Fyndanis, V., Economou, A., Protopapas, A. (2016). A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications. To appear in the Proceedings of the *10th Edition of the Language Resources and Evaluation Conference* 2016.