ΕΚΠΑΙΔΕΥΣΗ ΑΛΛΟΔΑΠΩΝ & ΠΑΛΙΝΝΟΣΤΟΥΝΤΩΝ ΜΑΘΗΤΩΝ

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ, ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Teaching Greek as a second language to young learners: The compilation of the Greek Learner Corpus

Despina Papadopoulou & Alexandros Tantos

Aristotle University of Thessaloniki

depapa@lit.auth.gr, atantos@gmail.com

Workshop on Language Education for Bilingual learners
Thessaloniki, 14 January 2012

Overview of the talk

- **Brief presentation of the Action "Educational & Linguistic Support for the teachers of foreigner & repatriate students"**
- **Sub-action: Compiling Greek Learner Corpus**
- **Learner Corpora and teaching Greek as a second/foreign language**
- **Error annotation scheme for GLC**
- **Implementation: Applying the Linguistic Annotation Framework [LAF] (ISO 24612) to GLC**

Aims of the action

To support the teachers of Greek as a second language in primary and secondary education by means of:

- a platform (Moodle, www.diapolis.auth.gr) for communication, networking, education material, asynchronous workshops, uploading teaching activities produced by the researchers and the teachers
- the organization of training workshops for teachers which aim at enhancing their linguistic awareness
- the production of education material based on CLIL, which has not been so far extensively employed in the Greek education system
- the generation of diagnostic tests as well as linguistic activities for testing pupils' linguistic competence (repetition, elicitation & comprehension tasks on agreement, determiners, verb morphology and prosody) and skills
- the compilation of the GLC based on written productions

Learner Corpora in Greek

Learner Corpora consist of students' written and oral productions and are annotated with respect to the students' errors. These errors are supposed to reflect their proficiency level and reveal aspects of their interlanguage. The methodology employed in the compilation of Learner Corpora is based on the computational analysis of linguistic data and is known as **computer-aided error analysis (CEA)**.

- 1st attempt to compile a Learner corpus in Greek by **Tzimokas (2010)**
- It consists of around **65,000 words** and **291 texts**
- This corpus is the first systematic attempt to codify a representative variety of adult learners' errors in Greek as a second/foreign language from an impressive number of L1 groups.
- **BUT:**
    - The error annotation scheme is complicated and inflexible for both groups of users, teachers of Greek as a second/foreign language and researchers.
    - It is based on a customary editing and validation tool with in-line annotated files in an XML output format, which is not compliant with any modern in-line XML-based linguistic annotation format (e.g., TEI Guidelines).

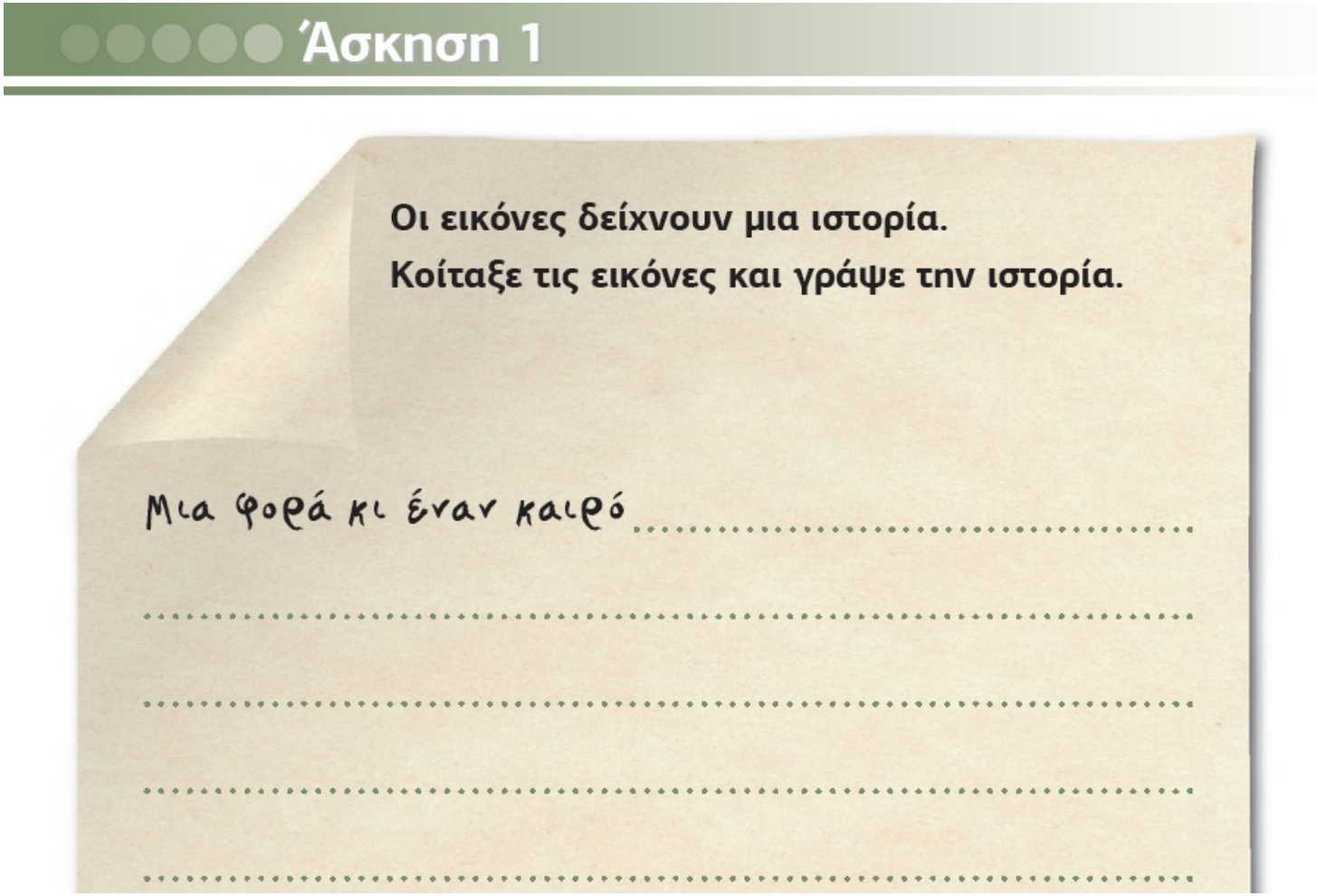
Greek Learner Corpus [GLC]: Data

The GLC consists of pupils' written productions within the diagnostic tests of proficiency in Greek (generated in the framework of our Action)

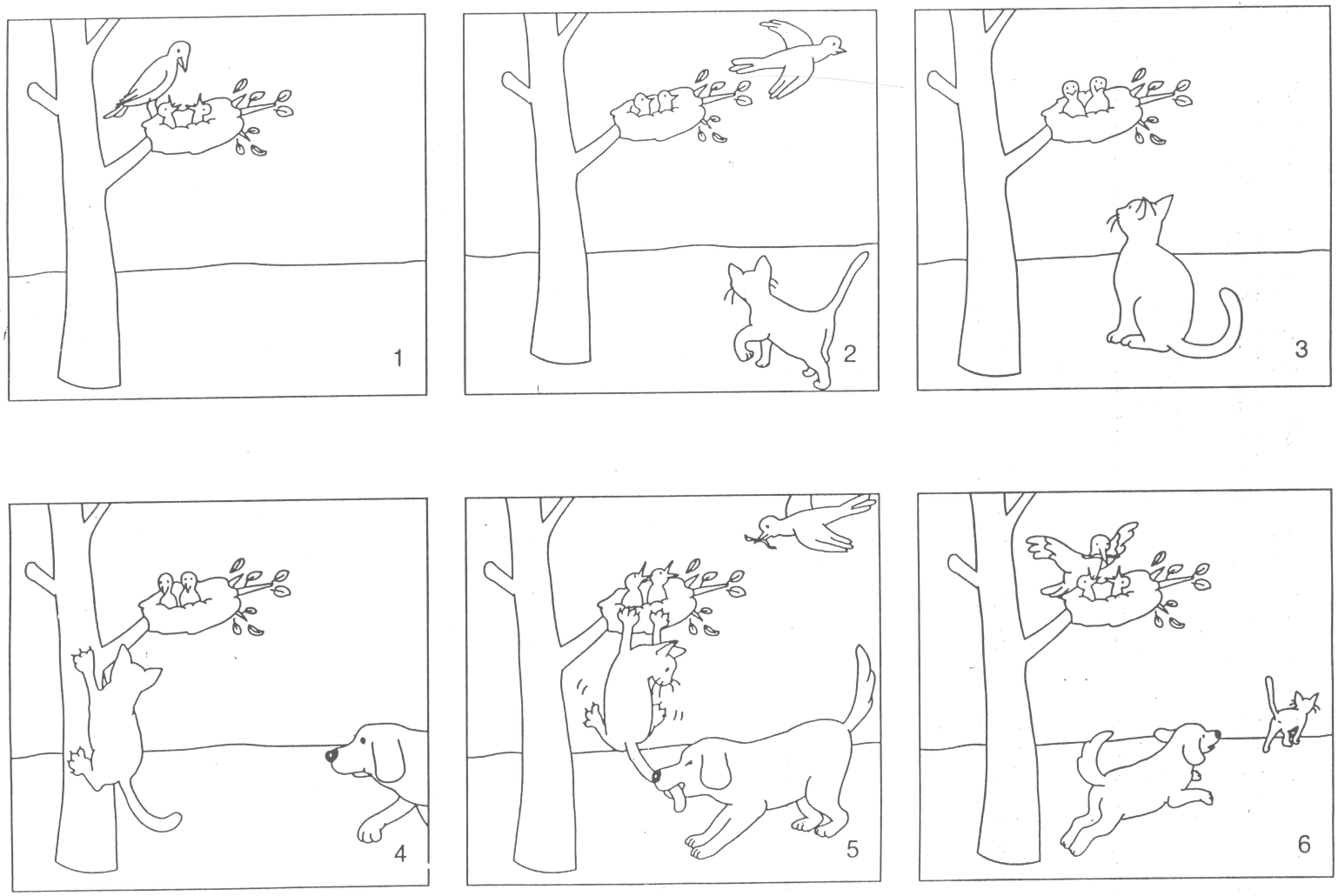
- 1000 tests collected so far, i.e. 2000 short texts
- Rich inventory of metadata, e.g. age, L1, age of onset, years of residence in Greece, years in the Greek education system, parents' ability in Greek

Greek Learner Corpus: Text data

Activity 1: common among Test I, Test II and Test III



The cat story



Greek Learner Corpus:

Text data

Activity 2: Test II

Χθες ήταν Κυριακή και πήγες μια εκδρομή με τους γονείς σου και με φίλους σου. Γράψε στο ημερολόγιό σου πώς πέρασες. Οι ερωτήσεις θα σε βοηθήσουν.

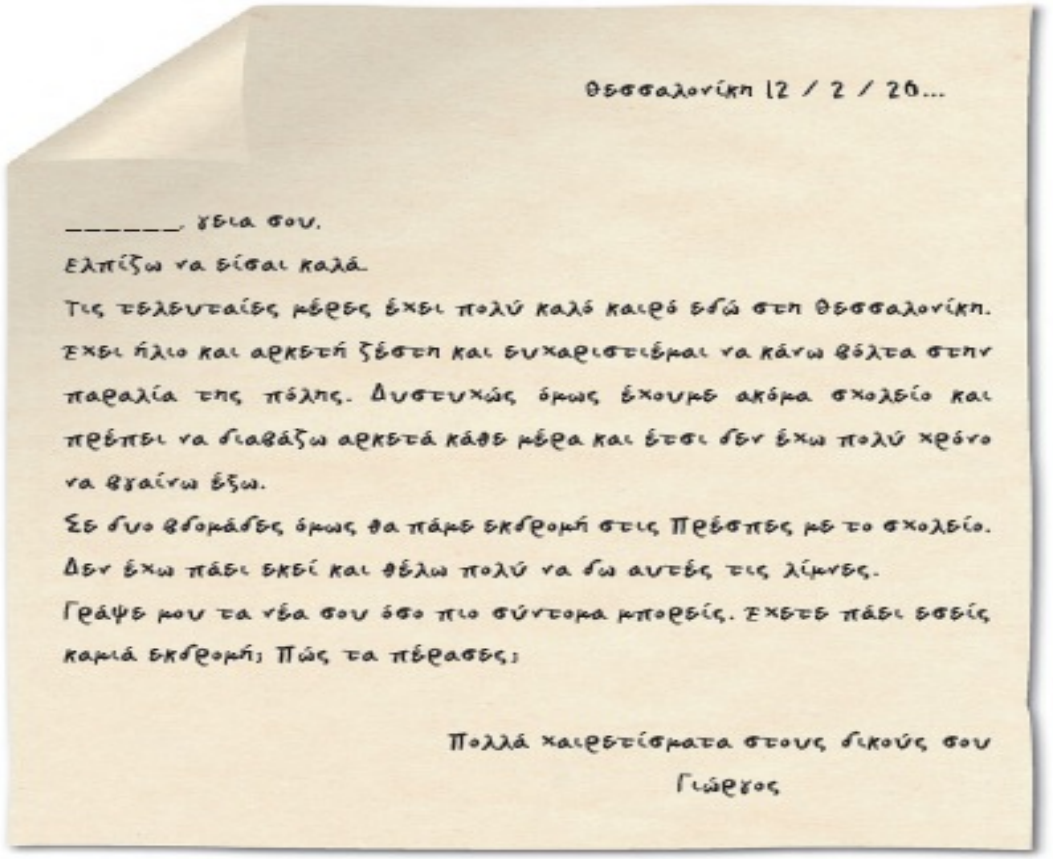
- **Με ποιους πήγατε εκδρομή;**
- **Πού πήγατε;**
- **Τι κάνατε στην εκδρομή;**
- **Τι θυμάσαι περισσότερο από την εκδρομή; Έγινε κάτι που σε έκανε να γελάσεις, να στενοχωρηθείς ή να φοβηθείς;**
- **Πώς τελείωσε αυτή η εκδρομή;**

Greek Learner Corpus:

Text data

Activity 2: Test III

Άσκηση 2



Θεσσαλονίκη 12 / 2 / 20...

_____, χθες σου.

Ελπίζω να είσαι καλά.

Τις τελευταίες μέρες έχει πολύ καλό καιρό εδώ στη Θεσσαλονίκη. Έχει ήλιο και αρκετή ζέστη και ευχαριστιέμαι να κάνω βόλτα στην παραλία της πόλης. Δυστυχώς όμως έχουμε ακόμα σχολείο και πρέπει να διαβάζω αρκετά κάθε μέρα και έτσι δεν έχω πολύ χρόνο να βγαίνω έξω.

Σε δυο βδομάδες όμως θα πάμε εκδρομή στις Πρέσπες με το σχολείο. Δεν έχω πάει εκεί και θέλω πολύ να δω αυτές τις λίμνες.

Γράψε μου τα νέα σου όσο πιο σύντομα μπορείς. Έχετε πάει εσείς καμιά εκδρομή; Πώς τα πέρασες;

Πολλά χαιρετίσματα στους δικούς σου

Γιώργος

Πριν από δύο μέρες πήρες το παραπάνω γράμμα από έναν φίλο σου που μένει στη Θεσσαλονίκη. Γράψε και εσύ ένα γράμμα για να του απαντήσεις, λέγοντας:

- **Αν έχεις πάει εκδρομή και πού πήγες.**
- **Πώς ήταν εκεί;**
- **Τι σου άρεσε στην εκδρομή; Γιατί;**
- **Τι δεν σου άρεσε; Γιατί;**

Error Annotation scheme of GLC

The error annotation scheme includes (a) parts of speech, (b) linguistic errors, e.g. determiners, clitics, tense, aspect etc, and (c) error category, e.g. omission, substitution, addition

Our error annotation scheme aims at thoroughly describing the pupils' errors avoiding to provide a theoretical interpretation of the errors.

Error Annotation scheme of GLC:

An example

(1) arxísame na *péksume
started.PERF.1PL SUBJ play.PERF.1PL

Error tag: _ASP_PERF

(2) *énas peristéri
a.**MASC**.SG.NOM pigeon.NEUT.SG.NOM

Error tag: (a) _AGR_GEN (b) _GEN

(3) *mía δéndra
a.**FEM**.SG.NOM tree.NEUT.PL.NOM

Error tag: (a) _GEN, (b) _AGR_GEN/NUM

Annotation Science

How do we encode linguistic data?

- Bracketings, XML tags, specialized format e.g. PropBank, ...

How do we encode ***overlapping alternative*** and ***conjunctive annotations***?

First Annotation Data Models:

- XCES = XML Corpus Encoding Standard
- Annotation Graphs [AG]

But:

- XCES was not comprehensive enough for many types of linguistic annotation
- AGs posed problems for representing hierarchical relations such as syntactic dependencies

Annotation Science

How do we handle diverse linguistic annotations of various resources for the same phenomena? (interoperability problem)

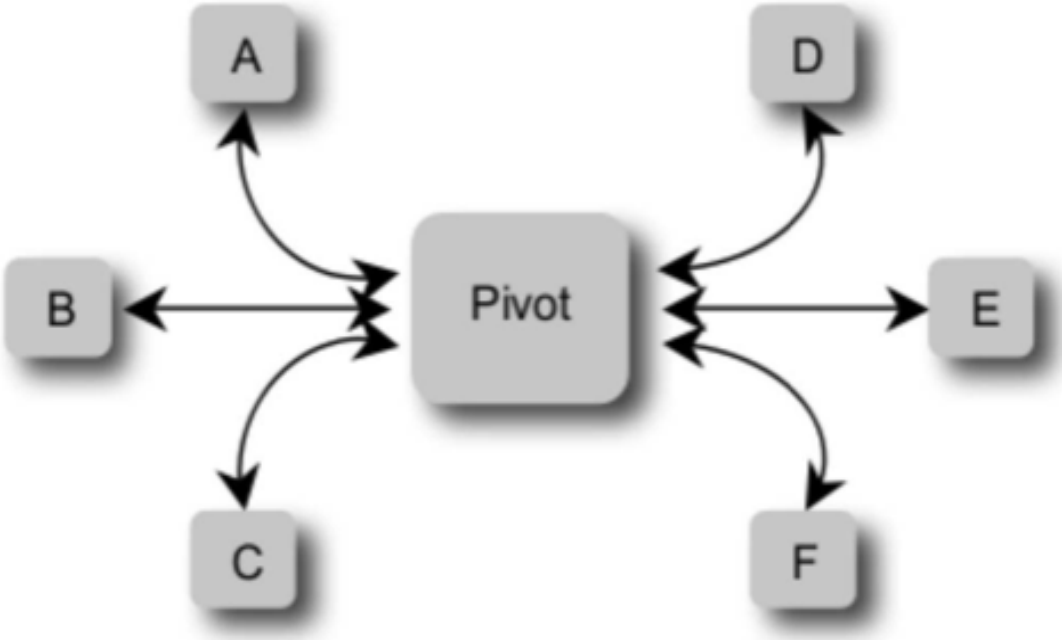
Corpora, Schemes, Formats

- Penn Treebank
- TIGER
- Negra
- TüBa-D/Z
- PropBank
- RST corpus
- Penn Discourse Treebank
- CALLHOME
- COCONUT
- EXMARaLDA
- MUC-7
- PAULA
- TUSNELDA
- XCES
- ...

Linguistic Annotation Framework:

The idea

An Interlingua for Annotation



Linguistic Annotation Framework [ISO 24612]

ISO 24612 covers inefficiencies of earlier annotation frameworks and aims at implementing some basic missing features in encoding, storing and processing of language resources:

- Descriptive adequacy for any level of linguistic annotation independently of theoretical preference or formatting/encoding constraints
- Independence of language production means (e.g. sound, voice, image)
- Interoperability, i.e. different software utilities have immediate access to any source of linguistically annotated data without discrepancies among them

LAF Data Model and Novelties in GLC

GLC adopts central architectural choices of LAF:

- *Stand-off Annotation* (i.e. Separating original data from annotation data in separate files) that allows for
  - **extensibility** of the resource in **ANY** kind of annotation
  - **multiple** error annotation efforts
  - **merging** of annotations and their **qualitative** and **quantitative** comparison
- *Separation of annotation content and annotation structure within a graph-based model*
  - **No formatting constraint** influences the structure of annotation (e.g. XML syntax of nesting elements) unlike what happens in standard in-line annotation frameworks (e.g. Text Encoding Initiative)

LAF Data Model and Novelties in GLC

GLC adopts central architectural choices of LAF:

- Primary data documents (read-only files)
- Base segmentation files (tokenization files)
- Any number of annotation documents (annotation links to tokens)
- Header documents (they contain meta-data; linguistically-relevant information about pupils' state)

GLC and GATE: Current and next steps

The platform General Architecture for Text Engineering (GATE) is the project's development platform: [DEMO]

GLC already collaborates with the team of American National Corpus (ANC) for councelling and standardization issues (LAF is in its final pre-published version).

GLC: Benefits within analysis Cycle

- **Improved classification** of error classes in various proficiency levels and thus one gets
  - a better understanding of what "moving from one proficiency level to a second" means
  - more refined classification criteria for learners and production of targeted activities (i.e., errors that remained unobservable are revealed within their context of use even if their frequency is not high though it exists.
  - and, practically, an improvement in the standardization of proficiency levels in Greek educational system

- Improved and more complicated **searching** of learners' errors by both teachers and researchers for both **qualitative** and **quantitative** analyses

- Production of improved **knowledge-rich tools of computer aided language learning** (e.g. spellcheckers, grammarcheckers) for both teachers and pupils

Many Thanks go to:
Katerina Aleksandri, Ifigenia Dosi, Konstandina Koutra, Katerina Meliadou & Katerina Pouliou

Also to our collaborators who visit the schools, collaborate with the teachers, collect the data, correct the tests and help in the training courses