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

Stand-off Annotation for Learner Corpora: Compiling Greek Learner Corpus (GLC)

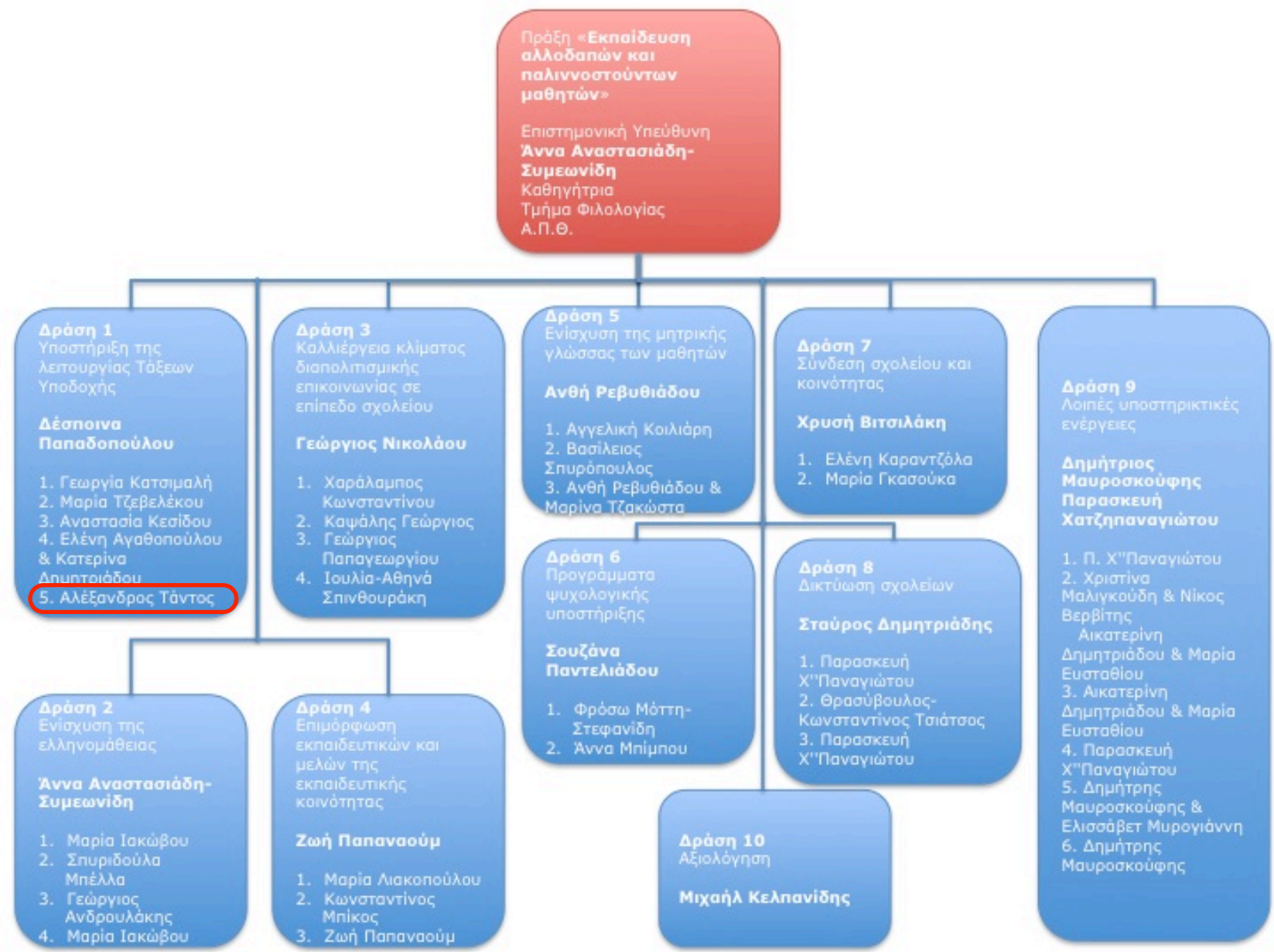
Alexandros Tantos & Despina Papadopoulou

Aristotle University of Thessaloniki
atantos@gmail.com, depapa@lit.auth.gr

Overview of the talk

- Brief presentation of the Action “Educational & Linguistic Support for Classes of Foreigner & Repatriate students”
- Greek Learner Corpus: Current Data
- Error Annotation in GLC
- Learner Corpora and Stand-off Annotation
- Linguistic Annotation Framework [LAF] (ISO 24612) and GLC
- Error annotation scheme for GLC
- Implementation: GATE and UAM Corpus Tool
- GLC: The analysis cycle

Educational and Linguistic Support for Classes of Foreign and Repatriate Students



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).

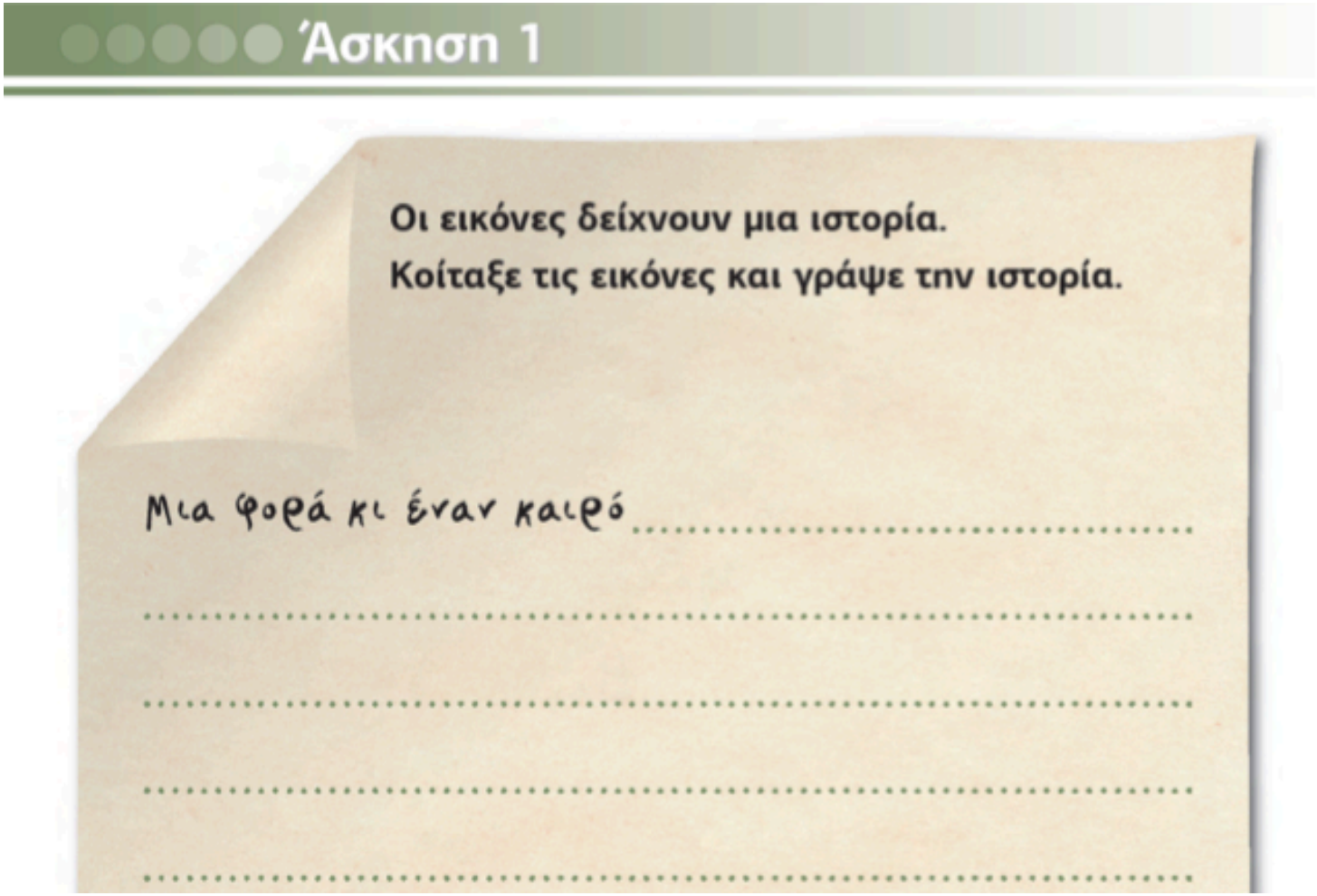
Greek Learner Corpus [GLC]: Data

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

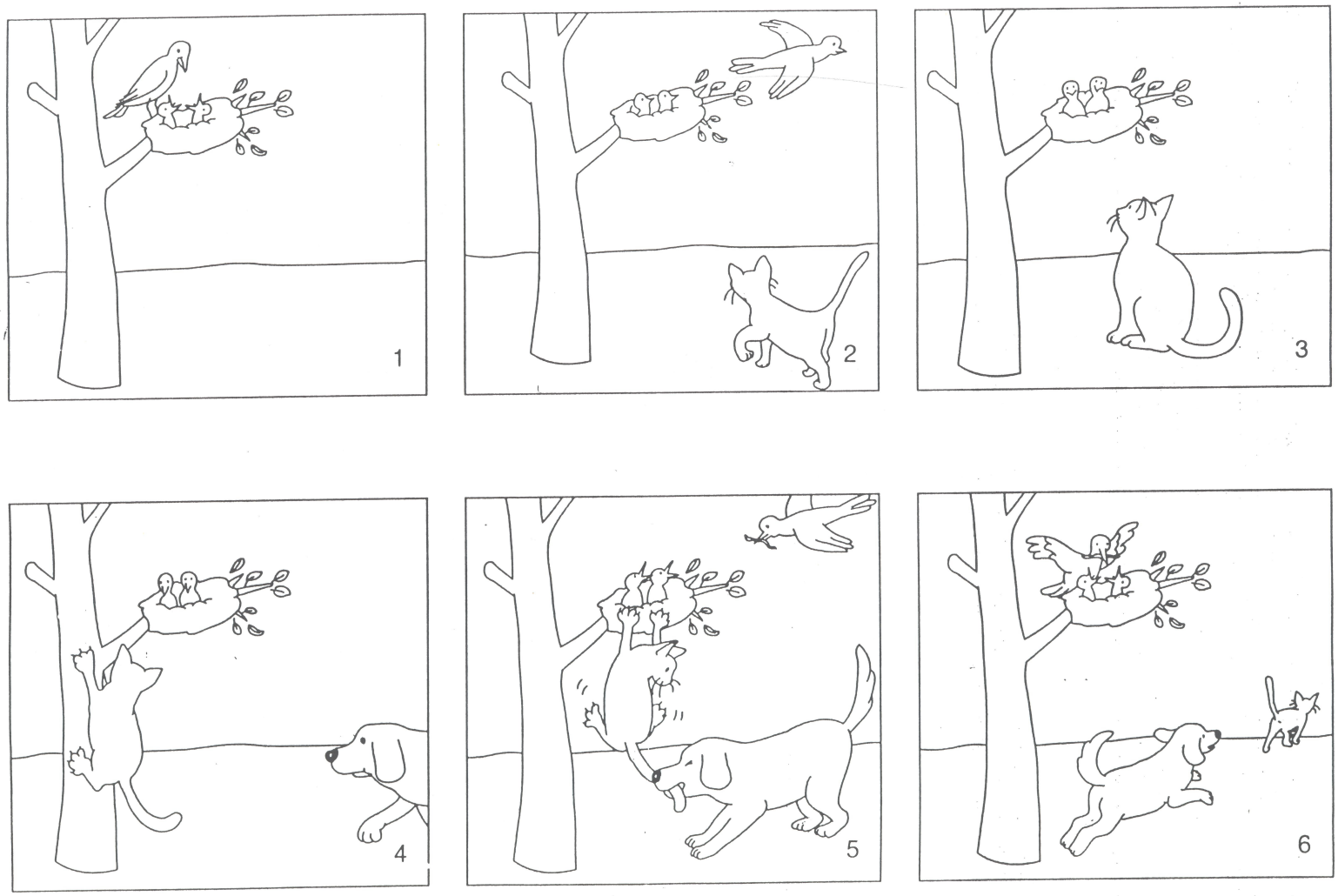
- 1000 tests collected so far, i.e. 2000 short texts / 1000 tests more to come by the end of June
- Rich inventory of metadata, e.g. age, L1, age of onset, years of residence in Greece, years in the Greek education system, parents' ability in Greek

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.

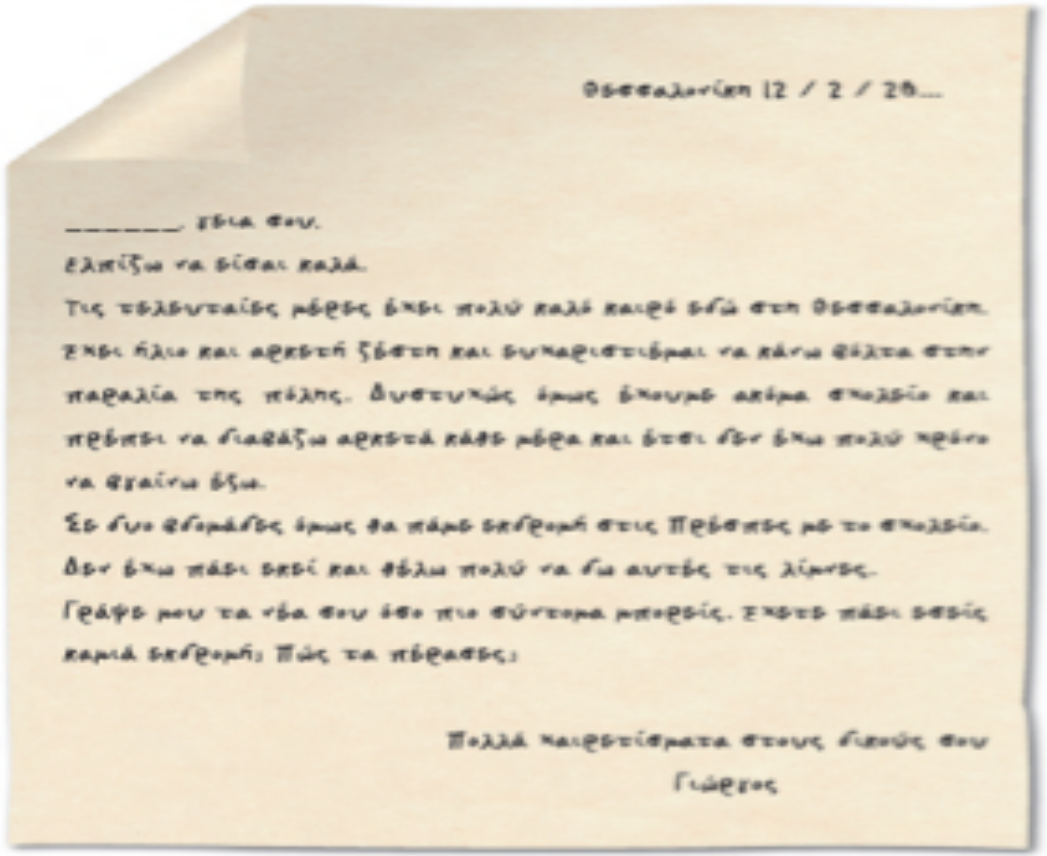
- Με ποιους πήγατε εκδρομή;
- Πού πήγατε;
- Τι κάνατε στην εκδρομή;
- Τι θυμάσαι περισσότερο από την εκδρομή; Έγινε κάτι που σε έκανε να γελάσεις, να στενοχωρηθείς ή να φοβηθείς;
- Πώς τελείωσε αυτή η εκδρομή;

Greek Learner Corpus:

Text data

Activity 2: Test III

Άσκηση 2



Θεσσαλονίκη 12 / 2 / 20...

_____, γειά σου.

Ελπίζω να είσαι καλά.

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

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

Γράψε μου τα νέα σου όσο πιο σύντομα μπορείς. Έχεις 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 Schemes

Error annotation schemas function as validation schemas and they have two purposes:

1) to delimit the number and kind of errors that we think are necessary to be encoded

2) to direct

a) *annotators* through XML validation schemas in various languages (such as XML Schemas, DTD, NG Schemas) in order to avoid common typos and/or more important mistakes in the process of annotation and

b) *machines* to interpret tag-positioning.

Error Annotation Schemes

The implementation of error taxonomies is driven by two kinds of belief:

1) of the linguists as to what the real needs are for designing a repository of errors

2) of other users that the error taxonomy should be manageable and flexible to cover *all* of their queries

Error Annotation Schemes and Correction

Many LCs include a correction of the learner's error in parallel to error annotation. However, error annotation and assignment of correct output are **only partly** identified.

Critical factors for the assignment of correct and intended output are:

1) L1 of the speaker-learner

2) level of proficiency of the speaker

3) reoccurrence of the same or similar mistakes

4) typographical clarity of learner's output

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: choices, characteristics

GLC:

- is designed in both a linguistically- and user-oriented way

- does not include corrected output as part of its annotation

- ambitions to encode errors as well as to annotate the rest of the words within an extra POS tagset that reflects the alignment between expected acquired phenomena and the categories used based on the information about the current state of the learner, coming from metadata

Error Annotation in GLC:

A simple case

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

Error tag: _ASP_PERF

Our error annotation scheme includes a rich inventory to include categories such as that the aspect of the infinitival should not be perfective. On the other hand, our error annotation is not done based on an explicit and detailed description of all possible grammatical categories mentioned in a grammar book.

Error Annotation in GLC:

An example

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

Error tag: (a) _AGR_GEN (b) _GEN

Two possible error sources from two different words:

Either in the agreement of the two words or in the assignment of gender to “the pigeon”. The one choice excludes the other, but both are possible with scope two different words.

Points of difficulty:

how to encode two possible overlapping errors (different offsets) with exclusion relationship,

how to query cases like this in the corpus

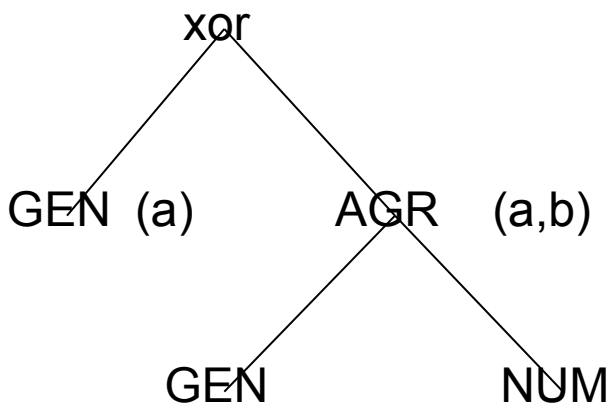
Error Annotation scheme of GLC:

An example

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

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

Two possible error sources from two different words with an alternative in the second disjunct:



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** that need to be taken into consideration in users' queries?

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, Formats, Schemes

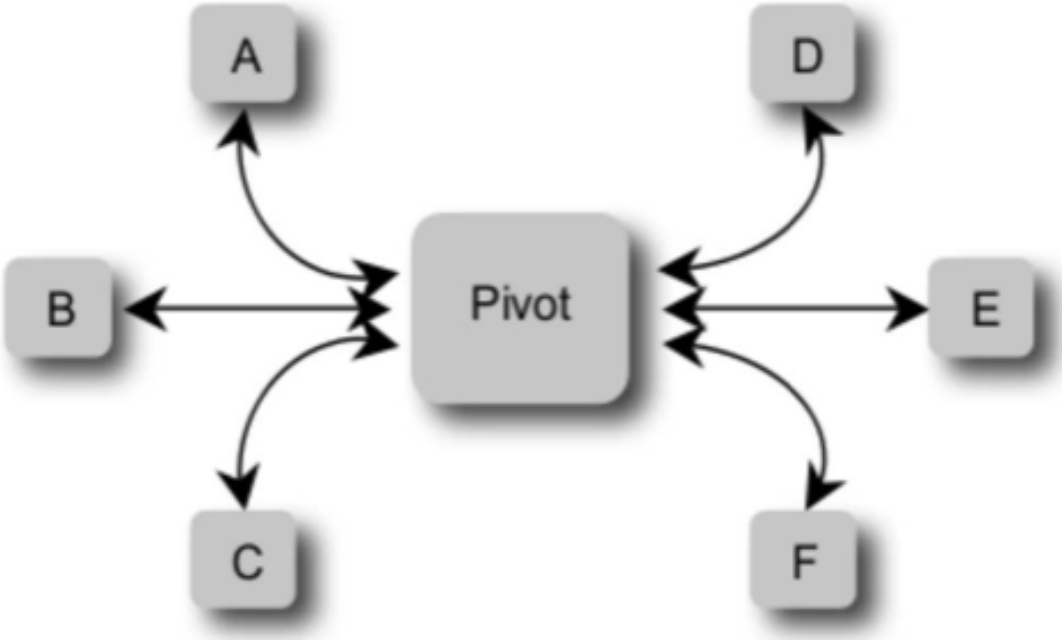
- Penn Treebank
- Penn Discourse Treebank
- Negra
- TueBA
- TIGER
- ...

- XCES
- MUC-7
- PAULA
- COCONUT
- ...

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)

LAF Data Model and Novelties in GLC

The linking property of stand-off annotation provided by LAF provides the chance to combine real data with both:

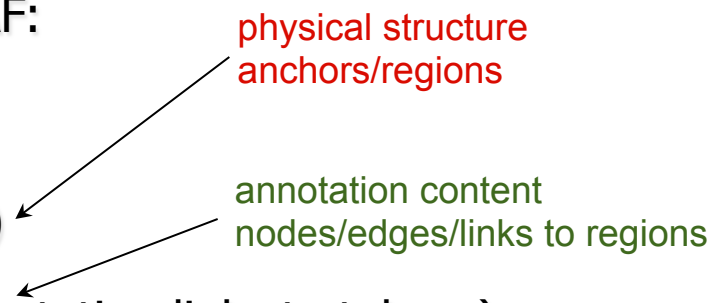
- the error annotation scheme and
- the GLC part of speech scheme in different layers avoiding encoding problems with overlapping offsets
- Different linking ways between nodes and between edges express the desired annotation dependencies

LAF data model is implemented within GraF's dump format, an XML Serialization that implements a graph-based annotation

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)



Annotation objects (content objects) are separate from reference points (regions)

LAF's Physical Structure (PS) in GLC

How is PS encoded?

- multiple way of referring to base or primary data (segmentation files have their own reference structure) / great scope of granularity of data representation
  - continuous segments
  - super and sub-segments

  (especially suitable for errors of merging words, e.g. "αποτιν" that merges "απο" + "τιν" can be reanalyzed, since there is another orthographic error of the determiner "τιν" --> "την")
  - discontinuous segments

  (often used in our LC for encoding badly-spelled or even words, such as "παρα θυ ρο" instead of "παράθυρο")
  - landmarks

  (single points, useful for a number of cases, e.g. encoding wrong presence or absence of accent markers in Greek)

LAF's Annotation Content (AC) in GLC

How is AC encoded?

Annotation content consists of two parts:

- a graph structure (includes nodes, edges, links to regions)
  - Nodes are connected to other nodes via edges as well as to reference points of the whole range of regions in the segmentation files
- an annotation structure (represents linguistic information / error annotation in our case)

Annotation objects (content objects) are attached to nodes/edges of the data

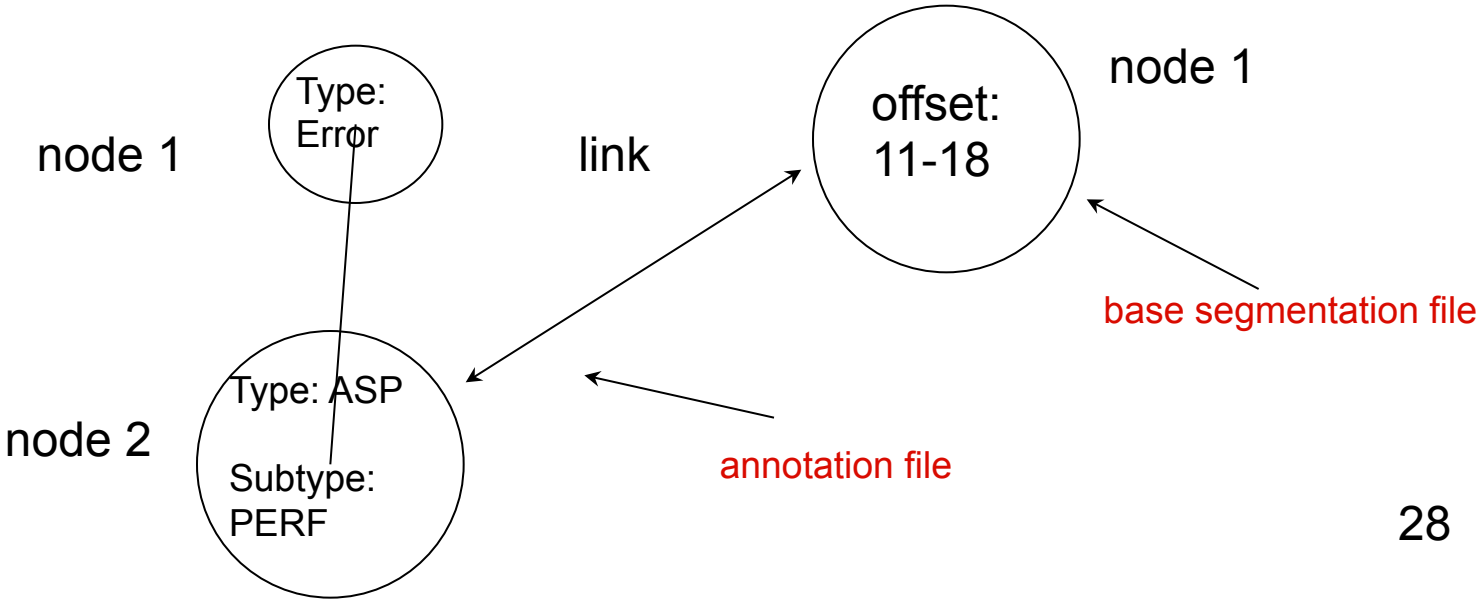
LAF Data Model and Novelties in GLC

Each annotation document consists of a "graph structure"

- "An annotation is defined as a label and a feature structure that is associated with a node or an edge in the graph"
- Nodes of the graph are NOT annotations, but only carriers of annotation.

(1) **0**a**1**r**2**x**3**í**4**s**5**a**6**m**7**e**8** **9**n**10**a**11** *p**12**é**13**k**14**s**15**u**16**m**17**e**18**
started.PERF.1PL SUBJ play.PERF.1PL

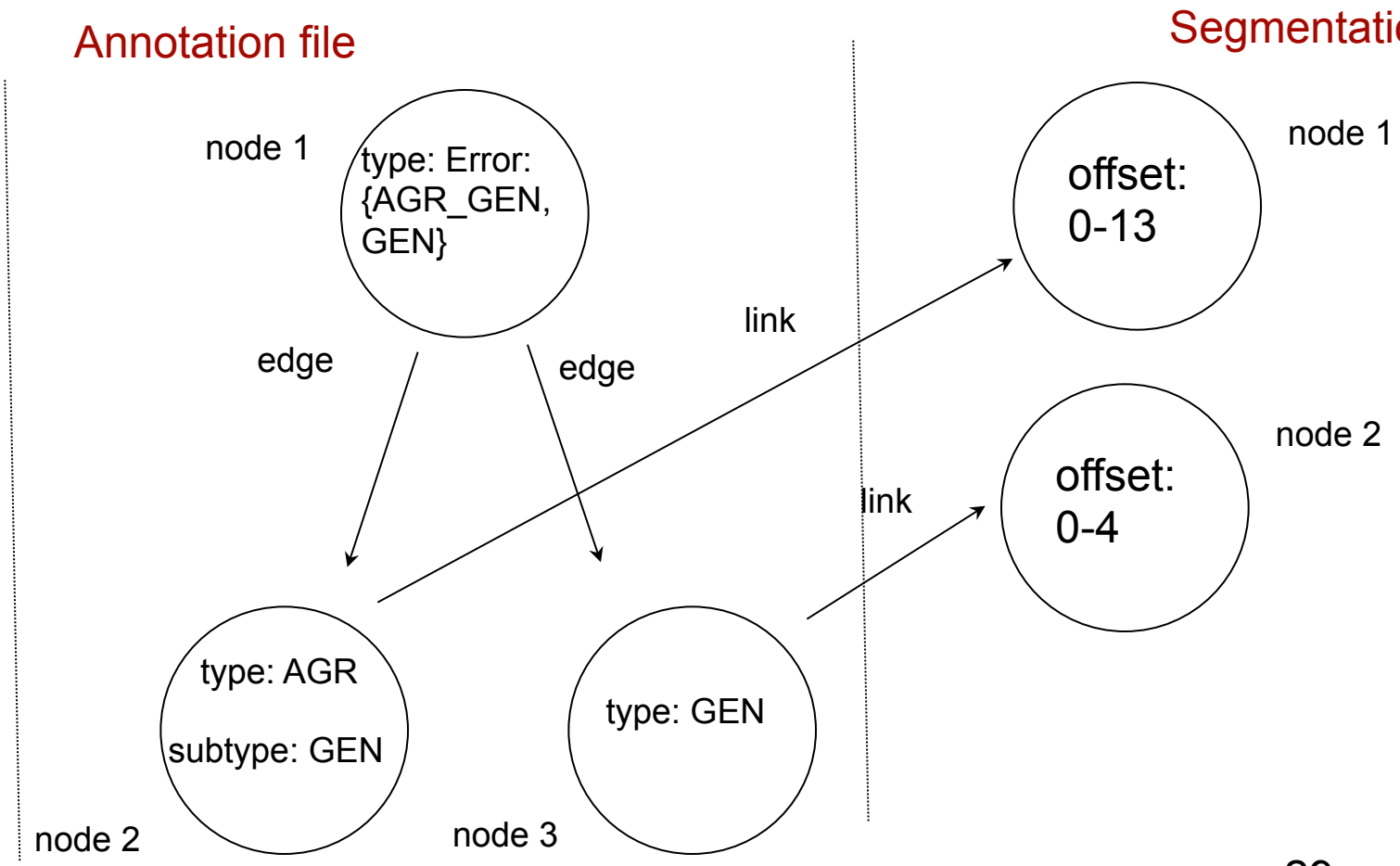
Error tag: _ASP_PERF



LAF Data Model and Novelties in GLC

(2) *0é1n2a3s4 p5e6r7i8s9t10é11r12i13
a.MASC.SG.NOM pigeon.NEUT.SG.NOM

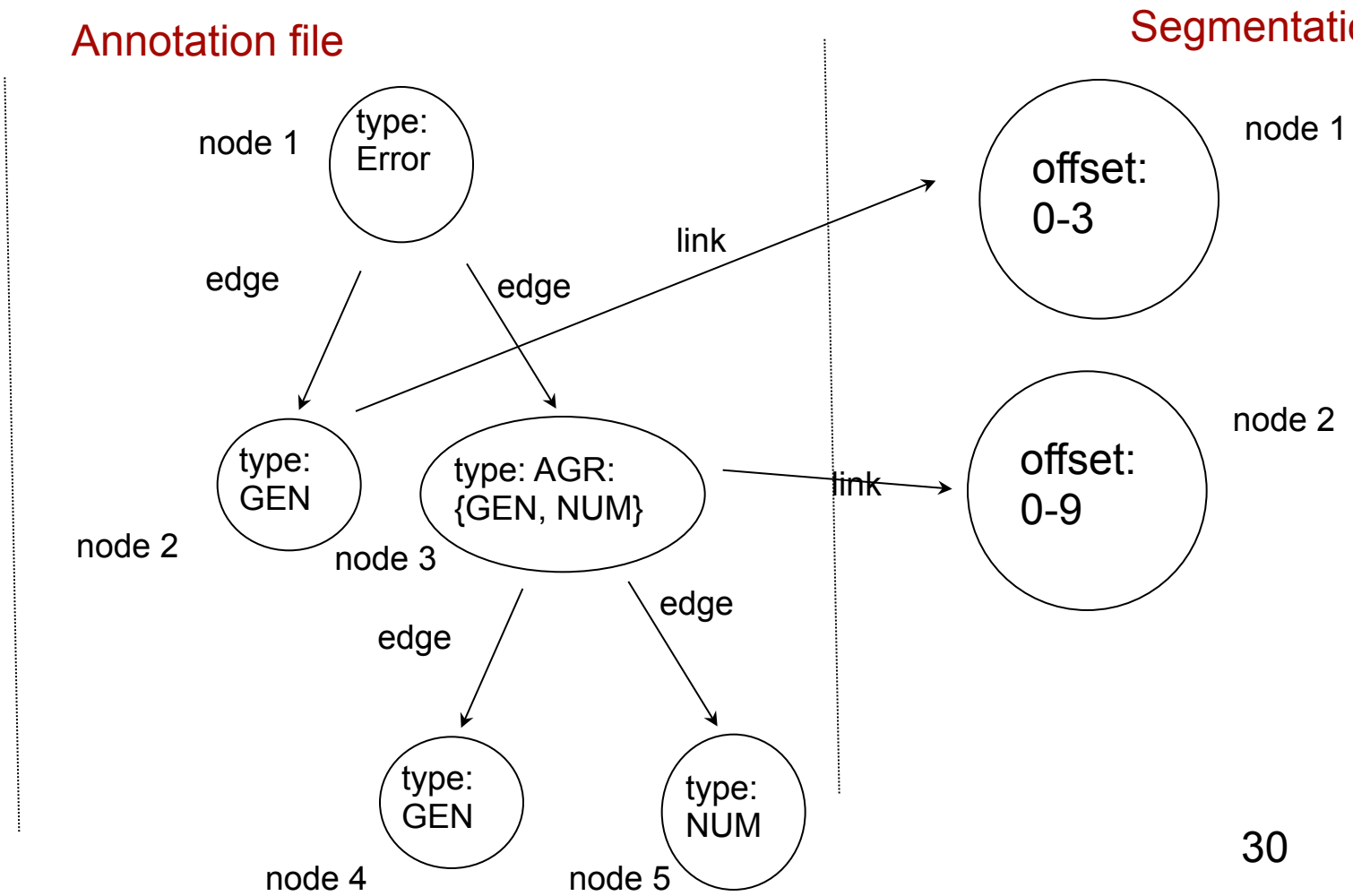
Error tag: (a) _AGR_GEN (b) _GEN



LAF Data Model and Novelties in GLC

(3) *0m1í2a3 δ 4é5n6d7r8a9
a.FEM.SG.NOM tree.NEUT.PL.NOM

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



GLC, GATE and UAM Corpus Tool: Current and next steps

The platform General Architecture for Text Engineering (GATE) has been used till recently as the project's annotation 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).

UAM Corpus Tool is a powerful annotation tool, the perspective project's development platform: [DEMO]

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, grammar checkers) 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