ΕΚΠΑΙΔΕΥΣΗ ΑΛΛΟΔΑΠΩΝ & ΠΑΛΙΝΝΟΣΤΟΥΝΤΩΝ ΜΑΘΗΤΩΝ

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

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

Ευρωπαϊκή Ένωση Ελληνική Δημοκρατία
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Επιχειρησιακό Πρόγραμμα Εκπαίδευση και Διά Βίου Μάθηση (ΕΣΠΑ 2007-2013)

Πράξη: «Εκπαίδευση Αλλοδαπών και Παλιννοστούντων Μαθητών»

Κωδικός Έργου: 85705

Άξονας Προτεραιότητας 1: «Αναβάθμιση της ποιότητας της εκπαίδευσης και προώθηση της κοινωνικής ενσωμάτωσης στις 8 Περιφέρειες Σύγκλισης»

Παραδοτέο:

1.5.1 ΕΠΙΣΗΜΕΙΩΤΗΣ ΣΩΜΑΤΩΝ ΚΕΙΜΕΝΩΝ ΜΑΘΗΤΩΝ

ΔΡΑΣΗ 1: ΥΠΟΣΤΗΡΙΞΗ ΤΗΣ ΛΕΙΤΟΥΡΓΙΑΣ ΤΩΝ ΤΑΞΕΩΝ ΥΠΟΔΟΧΗΣ

Υποδράση 1.5: ΨΗΦΙΟΠΟΙΗΣΗ ΓΡΑΠΤΩΝ & ΠΡΟΦΟΡΙΚΩΝ ΠΑΡΑΓΩΓΩΝ ΜΑΘΗΤΩΝ ΣΤΙΣ ΤΑΞΕΙΣ ΥΠΟΔΟΧΗΣ

Ομάδα έργου

ΠΑΠΑΔΟΠΟΥΛΟΥ ΔΕΣΠΟΙΝΑ	ΕΠΙΣΤΗΜΟΝΙΚΗ ΥΠΕΥΘΥΝΗ ΔΡΑΣΗΣ 1 ΕΠΙΚΟΥΡΗ ΚΑΘΗΓΗΤΡΙΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΓΛΩΣΣΟΛΟΓΙΑΣ, ΤΜΗΜΑ ΦΙΛΟΛΟΓΙΑΣ, ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΑΝΤΟΣ ΑΛΕΞΑΝΔΡΟΣ	ΕΠΙΣΤΗΜΟΝΙΚΟΣ ΥΠΕΥΘΥΝΟΣ ΥΠΟΔΡΑΣΗΣ 1.5 ΔΙΔΑΚΤΩΡ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΚΕΙΜΕΝΙΚΗΣ ΣΗΜΑΣΙΟΛΟΓΙΑΣ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΚΩΝΣΤΑΝΤΙΑΣ, ΓΕΡΜΑΝΙΑ ΜΕΤΑΔΙΔΑΚΤΟΡΙΚΟΣ ΕΡΕΥΝΗΤΗΣ Α.Π.Θ.
ΑΛΕΞΑΝΔΡΗ ΑΙΚΑΤΕΡΙΝΗ	ΔΙΔΑΚΤΟΡΙΚΗ ΦΟΙΤΗΤΡΙΑ ΓΛΩΣΣΟΛΟΓΙΑΣ ΤΟΥ ΤΜΗΜΑΤΟΣ ΦΙΛΟΛΟΓΙΑΣ, ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΔΟΣΗ ΙΦΙΓΕΝΕΙΑ	ΜΕΤΑΠΤΥΧΙΑΚΗ ΦΟΙΤΗΤΡΙΑ ΓΛΩΣΣΟΛΟΓΙΑΣ – ΔΙΔΑΚΤΙΚΗΣ ΤΟΥ ΤΜΗΜΑΤΟΣ ΓΕΡΜΑΝΙΚΗΣ ΓΛΩΣΣΑΣ & ΦΙΛΟΛΟΓΙΑΣ
ΚΟΥΤΡΑ ΚΩΣΤΑΝΤΙΑ	ΠΡΟΠΤΥΧΙΑΚΗ ΦΟΙΤΗΤΡΙΑ ΓΛΩΣΣΟΛΟΓΙΑΣ ΤΟΥ ΤΜΗΜΑΤΟΣ ΦΙΛΟΛΟΓΙΑΣ, ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΜΕΛΙΑΔΟΥ ΑΙΚΑΤΕΡΙΝΗ	ΠΡΟΠΤΥΧΙΑΚΗ ΦΟΙΤΗΤΡΙΑ ΓΛΩΣΣΟΛΟΓΙΑΣ ΤΟΥ ΤΜΗΜΑΤΟΣ ΦΙΛΟΛΟΓΙΑΣ, ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΠΟΥΛΙΟΥ ΑΙΚΑΤΕΡΙΝΗ	ΤΕΛΕΙΟΦΟΙΤΟΣ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΓΛΩΣΣΟΛΟΓΙΑ ΤΟΥ ΤΜΗΜΑΤΟΣ ΦΙΛΟΛΟΓΙΑΣ, ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Πίνακας καταγραφής εκδόσεων

Έκδοση	Ημερομηνία	Διαφορές από προηγούμενη έκδοση
ΤΕΛΙΚΗ	5/10/2011	-

Επιτελική σύνοψη

Σε αυτό το παραδοτέο παρουσιάζονται και αναλύονται τα ακόλουθα θέματα:

(α) βιβλιογραφική επισκόπηση σχετικά με τα Σώματα Κειμένων Μαθητών,

(β) την ψηφιοποίηση και δημιουργία του σώματος κειμένων μαθητών με την Ελληνική ως δεύτερη γλώσσα (ΣΚΕΙΜΑΘ),

(γ) το πλαίσιο επισημείωσης λαθών του ΣΚΕΙΜΑΘ,

(δ) πρακτικές επισημείωσης του ΣΚΕΙΜΑΘ,

(ε) το περιβάλλον επισημείωσης και χρήσης του ΣΚΕΙΜΑΘ.

Περιεχόμενα

Κατάλογος γραφημάτων

Κατάλογος πινάκων

1 Εισαγωγή

1.1 Στόχος Υποδράσης 1.5

Η υποδράση 1.5 στοχεύει στην ψηφιοποίηση γραπτών παραγωγών των μαθητών των τάξεων υποδοχής που προκύπτουν από τις δραστηριότητες παραγωγής γραπτού λόγου των διαγνωστικών τεστ ελληνομάθειας «Ας Μιλήσουμε Ελληνικά I, II & III) (για τα επίπεδα A1, A2, B1 & B2). Η ψηφιοποίηση ενός μεγάλου αριθμού παραγωγών αποσκοπεί στην χρηστική εκμετάλλευση του υλικού αυτού από δύο ομάδες χρηστών, τους δασκάλους των τάξεων υποδοχής και τους ερευνητές της ελληνικής ως δεύτερης γλώσσας. Από τη μια μεριά, οι δάσκαλοι των τάξεων υποδοχής θα μπορέσουν να αποκτήσουν μια συνολική θεώρηση της διαγλώσσας των μαθητών και να συστηματοποιήσουν τις διδακτικές τους παρεμβάσεις, όπου αυτό είναι αναγκαίο. Από την άλλη, οι ερευνητές της ελληνικής ως δεύτερης, μπορούν να αξιοποιήσουν το παραγόμενο υλικό για την ταξινόμηση των λαθών των μαθητών με βάση γραμματικά χαρακτηριστικά είτε ως προς τη μητρική γλώσσα των μαθητών είτε ως προς το επίπεδο γλωσσομάθειάς τους ή ακόμα και ως προς τη διαχρονική εξέλιξη της διαγλώσσας των μαθητών, καθώς προβλέπεται να συγκεντρωθούν γραπτές παραγωγές από τους ίδιους μαθητές σε διαφορετικές χρονικές περιόδους κατά τη διάρκεια του σχολικού έτους.

Η διαδικασία ψηφιοποίησης μπορεί να οδηγήσει σε μια βάση δεδομένων με γραπτές παραγωγές μαθητών, οι οποίες στη συνέχεια μπορούν να ταξινομηθούν με κριτήρια το σχολείο, την τάξη, την ηλικία των μαθητών και την ηλικία πρώτης ουσιαστικής έκθεσης στη δεύτερη γλώσσα ανάμεσα σε άλλα.[1] Αν και μια τέτοια βάση δεδομένων είναι χρηστική για τις δυο ομάδες χρηστών που προαναφέρθηκαν, έχει ωστόσο δύο κύρια μειονεκτήματα:

1. έχει περιορισμένη εμβέλεια ως προς την κατάδειξη των αδυναμιών των μαθητών και
2. δεν αποτελεί βάση για την ποσοτικοποίηση και συστηματοποίηση των λαθών των μαθητών και, επομένως, δεν αποτελεί αξιόπιστο εργαλείο για την ένδειξη του επιπέδου γλωσσομάθειας των μαθητών.

Για την όσο το δυνατό πιο λειτουργική αξιοποίηση του υλικού από τις δύο ομάδες χρηστών, δασκάλους και ερευνητές, και την ποικίλη αξιοποίησή του σε βάθος χρόνου, η υποδράση 1.5 καταφεύγει στη δημιουργία ενός σώματος κειμένων μαθητών της ελληνικής γλώσσας (ΣΚΕΙΜΑΘ). Αντλώντας μεθοδολογία και σειρά εργαλείων από δύο καθιερωμένα υποαντικείμενα της γλωσσολογίας, τη γλωσσολογία σωμάτων κειμένων (ΓΣΚ) και την υπολογιστική γλωσσολογία

[1] Ακόμη και η πιστή αντιγραφή από το αυθεντικό γραπτό υλικό δεν αποδεικνύεται εύκολη υπόθεση λόγω του δυσερμήνευτου γραφικού χαρακτήρα των μαθητών. Αυτό διαπιστώνουμε καθημερινά στο πρώτο στάδιο του προγράμματός μας όπως αναφέρεται και στην ενότητα 3.1.

(ΥΓ), το ΣΚΕΙΜΑΘ είναι μια συλλογή κειμένων των μαθητών που επισημειώνεται με μια προσυμφωνημένη και καλά σχεδιασμένη ιεράρχηση των λαθών των μαθητών. Ο σχεδιασμός και η υλοποίηση του ΣΚΕΙΜΑΘ αποτελεί μια χρονοβόρα διαδικασία που απαιτεί τη συνεργασία μιας ομάδας γλωσσολόγων με καλή γνώση της γραμματικής της ελληνικής καθώς και των γλωσσολογικών αρχών που τη διέπουν, αλλά και γνώση των εργαλείων εκείνων που θα μετατρέψουν τα επισημειωμένα κείμενα σε αναπόσπαστα εργαλεία υποβοήθησης της διδασκαλίας της ελληνικής και στοχευμένης έρευνας στην κατάκτηση της ελληνικής ως δεύτερης γλώσσας.

1.2 Δομή Παραδοτέου

Για την αποσαφήνιση της δουλειάς που διεξάγεται στα πλαίσια της υποδράσης 1.5, το παρακάτω κείμενο δομείται ως εξής:

Η ενότητα δύο είναι αφιερωμένη σε μια σύντομη αλλά περιεκτική βιβλιογραφική επισκόπηση σε σχέση με τα σώματα κειμένων μαθητών, προκειμένου να γίνει κατανοητή και να αναδειχθεί η σημασία και η κατεύθυνση των σωμάτων κειμένων για το πρόγραμμά μας.

Η ενότητα τρία περιγράφει λεπτομερειακά τη διαδικασία σχεδίασης και υλοποίησης του ΣΚΕΙΜΑΘ μέχρι τώρα. Ξεκινώντας με την περιγραφή της ψηφιοποίησης και αναφορά σε παραδείγματα δυσκολιών, ακολουθεί η περιγραφή του παρόντος πλαισίου λαθών επισημείωσης, το οποίο αποτελεί το πιο πολυσύνθετο, θεωρητικά, κομμάτι της υποδράσης 1.5 και το οποίο θα εξακολουθεί να διαμορφώνεται, να αλλάζει και να εξελίσσεται και το επόμενο διάστημα μέχρι το τέλος των εργασιών της υποδράσης. Ένας συγκεντρωτικός πίνακας περιέχει την ιεράρχηση των συμφωνημένων και επεξεργασμένων κατηγοριών λαθών μέσα από σειρά συναντήσεων της ομάδας της υποδράσης 1.5 τους τελευταίους μήνες, καθώς και τις συντομογραφικές συμβάσεις οι οποίες κωδικοποιούνται τεχνικά ως ετικέτες για τη γλώσσα σήμανσης XML (Extensible Mark-up Language). Πέρα από την ταξινόμηση και σύντομη περιγραφή των κατηγοριών λαθών, ο συγκεκριμένος πίνακας περιέχει και ενδεικτικά παραδείγματα που καταδεικνύουν αφενός την εμβέλεια χρήσης των ετικετών και αφετέρου τις δυσκολίες που τυχόν παρουσιάζονται κατά την επισημείωση.

Η υποενότητα 3.2 είναι το τεχνικό κομμάτι υλοποίησης του ΣΚΕΙΜΑΘ. Στην υποενότητα 3.2 περιγράφονται αναλυτικά οι τεχνικές δυσκολίες και επιλογές με στόχο τη βέλτιστη αξιοποίηση της θεωρητικής μελέτης σχετικά με τις κατηγορίες λαθών και την ευκολία χρήσης του ΣΚΕΙΜΑΘ από δασκάλους κι ερευνητές. Ειδικότερα, το ΣΚΕΙΜΑΘ στηρίζεται σε πρωτοποριακές στρατηγικές επισημείωσης, οι οποίες επιτρέπουν πολύπλευρη αξιοποίηση του επισημειωμένου υλικού, χωρίς να υπάρχει κίνδυνος αλλοίωσης των αυθεντικών κειμένων. Επιπλέον, η υποενότητα 3.2 περιγράφει το τεχνικό κομμάτι της διαδικασίας επισημείωσης από τους επισημειωτές της ομάδας, μέσα από ένα μικρό επισημειωμένο κείμενο. Το περιβάλλον επισημείωσης γίνεται στη βάση της πλατφόρμας επεξεργασίας φυσικού λόγου GATE [General Architecture for Text Engineering] και αποτελεί όχι μόνο ένα λειτουργικό μέσο για την επισημείωση λαθών που αποκρύπτει σε μεγάλο

βαθμό την πολυπλοκότητα της γλώσσας XML και εξοικονομεί χρόνο για την ομάδα των επισημειωτών, αλλά και με τις χρωματισμένες περιοχές του κειμένου, θα αποτελέσει και το διαδραστικό περιβάλλον χρήσης για τις δύο ομάδες, ερευνητές και δασκάλους.

2 Σώματα κειμένων μαθητών: Βιβλιογραφική επισκόπηση

Τα σώματα κειμένων μαθητών (ΣΚΜ) αποτελούνται από ψηφιοποιημένα κείμενα παραγωγών μαθητών σε μια ξένη/δεύτερη γλώσσα (Leech 1998) και έχουν αποδειχθεί χρήσιμο και σημαντικό εργαλείο για τα πεδία της εκμάθησης και της διδασκαλίας δεύτερης γλώσσας, καθώς παρέχουν μεγάλα δείγματα αυθεντικών γλωσσικών δεδομένων των μαθητών (Granger 1998, Leech 1998, Pravec 2002). Τα κείμενα αυτά ταξινομούνται με κριτήρια χρήσιμα και για τα δύο επιστημονικά πεδία (ηλικία, επίπεδο γλωσσομάθειας και πρώτη γλώσσα των μαθητών) και αποκαλύπτουν πτυχές της διαγλώσσας για τους γλωσσολόγους ερευνητές, καθώς και σημεία διδακτικής παρέμβασης και εστίασης για τον εκπαιδευτικό.

Το κύριο κοινό χαρακτηριστικό όλων των σύγχρονων ΣΚΜ είναι ότι παρέχουν επισημειωμένα κείμενα βασισμένα πάνω σε ένα προσυμφωνημένο σύστημα ταξινόμησης λαθών των μαθητών για τη γλώσσα στόχο. Μετά το AILA symposium το 1996 και το First International Symposium on Learner Corpora στο Hong Kong το 1999, όταν το μόνο ΣΚΜ υπήρξε το International Corpus of Learner English (ICLE) για την αγγλική, αναγνωρίστηκε άμεσα η σημασία των ΣΚΜ και μια ολόκληρη νέα ερευνητική κοινότητα κινήθηκε για τη δημιουργία ΣΚΜ. Στον πίνακα 1, οι Díaz-Negrillo και Fernández-Domínguez (2006) δίνουν μια λίστα με τα υπάρχοντα ΣΚΜ μόλις το 2002 και αποκαλύπτουν την αυξανόμενη ανάγκη για τη δημιουργία και αξιοποίηση ΣΚΜ μέσα σε μόλις τρία χρόνια. Σήμερα, ο αριθμός και η ποικιλία των ΣΚΜ σε όλο τον κόσμο συνεχώς αυξάνεται με σειρά συμποσίων και συνεδρίων να φιλοξενούν τις ερευνητικές ομάδες που τα δημιουργούν.

Learner Corpus	Approximate Size (in words)	L2 Level	L1	Purpose	Mode	Learner Language
CBACLE	1,000,000	Various	Chinese	Academic	Written	English
CLC	20,000,000	Various	Various	Commercial	Written	English
C–LEG	28,000	Advanced	English	Academic	Written	German
FALKO	36,000	Advanced	Unspecified	Academic	Written	German
FRIDA	200,000	Advanced	Various	Academic	Written	French
HKUST	25,000,000	Upper secondary education	Chinese	Academic	Written	English
ICLE	2,000,000	Advanced	Various	Academic	Written	English
JEFLL	700,000	Various	Japanese	Academic	Spoken and written	English
LLC	10,000,000	Various	Various	Commercial	Written	English
MELD	100,000	Advanced	Unspecified	Academic	Written	English
NICT JLE	2,000,000	Various	Japanese	Academic	Spoken	English
PELCRA	500,000	Various	Polish	Academic	Written	English

Πίνακας 1. Λίστα Σωμάτων Κειμένων Μαθητών

Η πλειονότητα των ΣΚΜ βασίζεται στη θεωρία της Αντιπαραβολικής Ανάλυσης της Διαγλώσσας [Contrastive Interlanguage Analysis (CIA)] (Granger 2008). Σε αντίθεση με την κλασική αντιπαραβολική ανάλυση, η οποία συγκρίνει διαφορετικές γλώσσες, η CIA συγκρίνει διαγλώσσες της ίδιας γλώσσας και εμπεριέχει δύο υποκατηγορίες σύγκρισης: (α) σύγκριση της διαγλώσσας με τη γλώσσα στόχο και (β) σύγκριση ανάμεσα σε διαφορετικές διαγλώσσες διαφορετικών ατόμων. Η πρώτη περίπτωση σύγκρισης παίζει σημαντικό ρόλο στο διαχωρισμό των ιδιοσυγκρασιακών χαρακτηριστικών της γλώσσας στόχου, ενώ η δεύτερη αξιολογεί το βαθμό της γενίκευσης των χαρακτηριστικών της διαγλώσσας σε συνάρτηση με τα άτομα που μαθαίνουν τη γλώσσα και τις γλωσσικές περιστάσεις (Granger 2008: 341). Η δεύτερη μορφή σύγκρισης δεν έχει δεχτεί κριτική από τους ειδικούς για την κατάκτηση της δεύτερης γλώσσας, σε αντίθεση με την πρώτη περίπτωση σύγκρισης, η οποία έχει θεωρηθεί «ένοχη» ότι οδηγεί σε «συγκριτική πλάνη» (πρβ. Bley-Vroman 1983), καθώς συγκρίνοντας τη διαγλώσσα με τη μητρική γλώσσα, η πρώτη αδυνατεί να εξεταστεί χωριστά.

Σε σχέση με την πρακτική αξιοποίηση των επισημειωμένων κειμένων, ο Leech (1998) και η Granger (1998, 2002) θέτουν ποικίλα ζητήματα αναφορικά με τη μεθοδολογία που ακολουθείται για τα ΣΚΜ. Η Meunier (1998) αναφέρεται πιο ειδικά στα εργαλεία και λογισμικά που μπορούν να χρησιμοποιηθούν στην έρευνα των ΣΚΜ, οι Van Rooy and Schöfer (2003) ασχολούνται με την αξιοπιστία της επισημείωσης των μερών του λόγου, ενώ ο de Mönnink (2000) εξετάζει την σκοπιμότητα της συντακτικής ανάλυσης (parsing) των ΣΚΜ. Άλλες περιγραφές της μεθοδολογίας για τη CIA υπάρχουν στους Granger (1996) και Gilquin (2001), ενώ οι αρχές της Ανάλυσης Λαθών Υποβοηθούμενης από Υπολογιστές [Computer-Aided Error Analysis, (CEA)] παρουσιάζονται από τους Milton και Chowdhury (1994), Dagneaux et al. (1998), de Haan (2000) και Nicholls (2003).

3 Επισημειωτής λαθών μαθητών

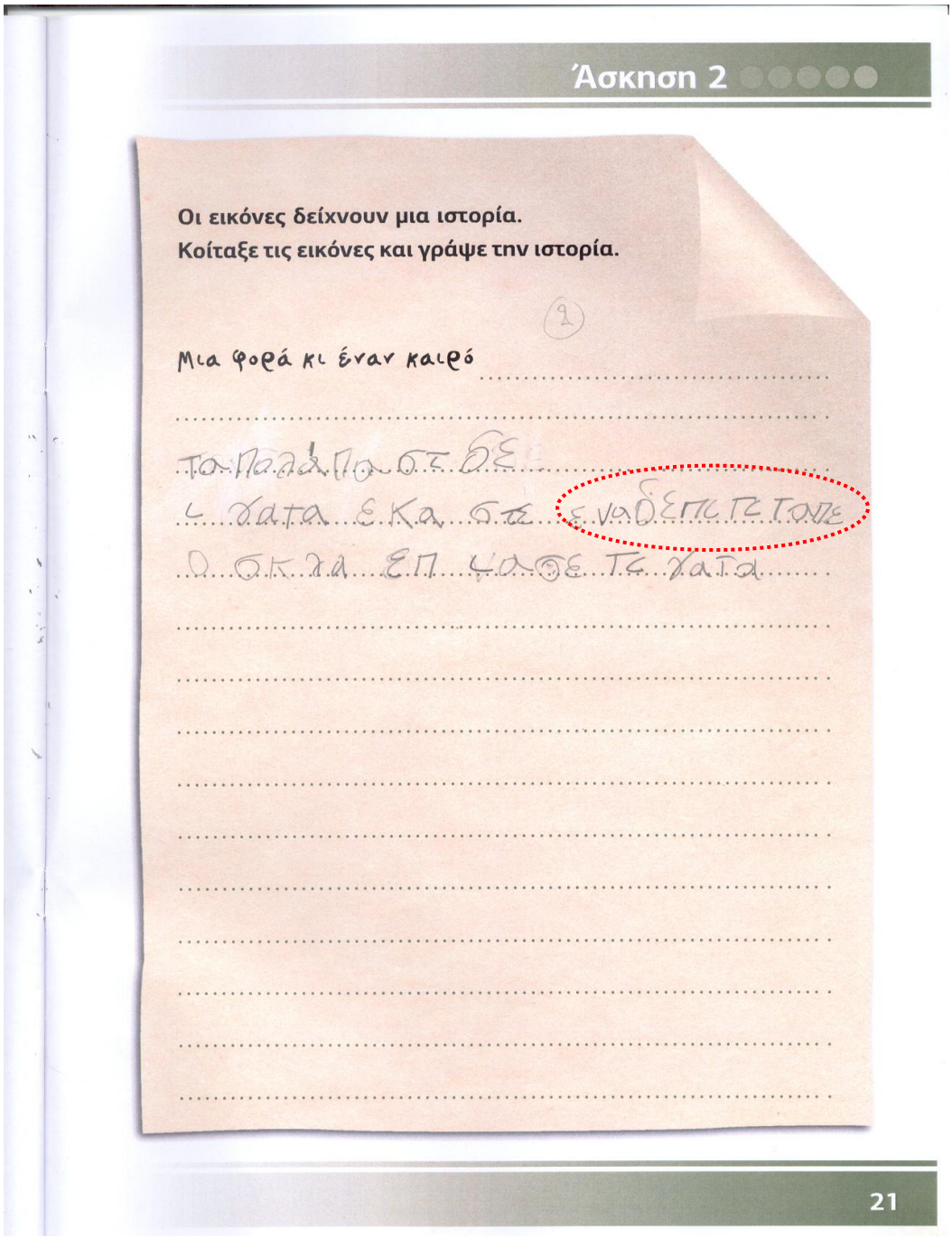
3.1 Ψηφιοποίηση γραπτών παραγωγών

Στη διαδικασία δημιουργίας του επισημειωτή λαθών μαθητών για το ΣΚΕΙΜΑΘ, η ψηφιοποίηση των γραπτών παραγωγών αποτελεί το πρώτο βήμα και πρόκληση για την κωδικοποίηση των λαθών των μαθητών. Σε κάθε στάδιο σε αυτή τη φάση της έρευνας προσπαθήσαμε να ψηφιοποιήσουμε τις γραπτές παραγωγές των μαθητών αποδίδοντάς τες όσο το δυνατόν πιο πιστά στο πρωτότυπο, αποφεύγοντας τον κίνδυνο της παρέμβασης στο κείμενο που συχνά αντιμετωπίζαμε. Η ψηφιοποίηση των 55 κειμενικών παραγωγών των μαθητών έχει πραγματοποιηθεί από δύο μέλη της ομάδας της υποδράσης 1.5 και αξίζει να σημειωθεί ότι η συμφωνία μεταξύ τους ως προς την ψηφιοποίηση των γραπτών παραγωγών δεν ήταν δεδομένη σε όλη τη διάρκεια της μεταγραφής των υπαρχόντων.

Η πρώτη δυσκολία που παρουσιάστηκε κατά την ψηφιοποίηση των κειμένων ήταν αρχικά η φωτοτυπημένη μορφή τους, που παρουσιάζει χαμηλή ποιότητα και ακρίβεια. Η σύγκριση, ωστόσο, των φωτοτυπημένων κειμένων με τα αυθεντικά περιόρισε τις δυσκολίες στην ψηφιοποίησή τους.

Μια άλλη δυσκολία που παρουσιάστηκε ήταν η αποκωδικοποίηση των γραφημάτων. Ενδεικτικά, μία δυσκολία αποκωδικοποίησης εμφανίστηκε στη φράση «ε ναδεπι τι τονζε» ενός κειμένου (βλ. Γράφημα 1). Η φράση αποκωδικοποιήθηκε με αυτό τον τρόπο μετά από εξονυχιστικό έλεγχο και λαμβάνοντας υπόψη τις εικόνες που δίνονται στους μαθητές για να περιγράψουν. Μια άλλη δυσκολία αποκωδικοποίησης που συναντήσαμε ήταν ότι τα κενά μεταξύ των λέξεων δεν ήταν πάντα σαφή (βλ. Γράφημα 1, 2 & 3), ενώ το πρόβλημα γινόταν εντονότερο όταν η γνώση των κενών ήταν άρρηκτα συνδεδεμένη με το νόημα του κειμένου (π.χ. τρωφειμα την ειδε ή τροφει μα την ειδε, βλ. Γράφημα 4). Επιπλέον, παρατηρούνται πολλά προβληματικά σημεία ως προς την αποσαφήνιση γραφημάτων. Για παράδειγμα, το γράφημα «σ» συμβολίζεται με έναν χαρακτήρα που μοιάζει με το γράφημα «ο», («οτα/οτο/οτη» αντί για «στα/στο/στη», «ουνέχεια» αντί για «συνέχεια», βλ. Γράφημα 2 & 3), το «αι» εμφανίζεται ως «ου» («παυδιά/παύζαμε/παυχνίδι», βλ. Γράφημα 3), του «α» με το «ο» και ταυτόχρονα του «ι» με το «υ» («πουχνίδια /πούζαμε»). Στις περιπτώσεις αυτές προσπαθήσαμε να δούμε πώς αποδίδει ο μαθητής το ίδιο γράφημα σε διαφορετικό περιβάλλον για να συμπεράνουμε αν πρόκειται για παραδρομή ή όχι. Εάν δεν υπήρχε αυτή η δυνατότητα, καθώς το αντιπροσωπευτικό δείγμα κειμένου από τον ίδιο μαθητή δεν ήταν επαρκές, γινόταν συμφωνία μεταξύ των επισημειωτών. Για περιπτώσεις όπου ανάμεσα στις συλλαβές της ίδιας λέξης υπήρχε μεγάλο κενό, αποφασίσαμε πως το τεμάχιο (chunk) αυτό θα επισημειώνεται ως μία λέξη (π.χ κι νι γο σε το επισημειώνουμε ως κινιγοσε). Τέλος, ο τόνος σε πολλές περιπτώσεις δεν ήταν ξεκάθαρος με συνέπεια να μην είμαστε σίγουροι αν έχει χρησιμοποιηθεί από τον μαθητή (βλ. Γράφημα 5). Για τέτοιες περιπτώσεις αποφασίσαμε κατά την επισημείωση να λαμβάνεται η λέξη ως σωστά τονισμένη.

Η διαδικασία διασαφήνισης των γραπτών παραγωγών των μαθητών αποτελεί το πρωταρχικό μέλημα της υποδράσης, προκειμένου να αποφευχθεί αθέμιτος επηρεασμός στην ανάλυση των λαθών σε επόμενο στάδιο. Η ψηφιοποίηση θα αναληφθεί από τους εκπαιδευτικούς των τάξεων υποδοχής και όποιες δυσκολίες προκύψουν στην ερμηνεία των γραφημάτων και των λέξεων θα γίνεται επί τόπου.



Άσκηση 2

Οι εικόνες δείχνουν μια ιστορία.
Κοίταξε τις εικόνες και γράψε την ιστορία.

Μια φορά κι έναν καιρό .....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

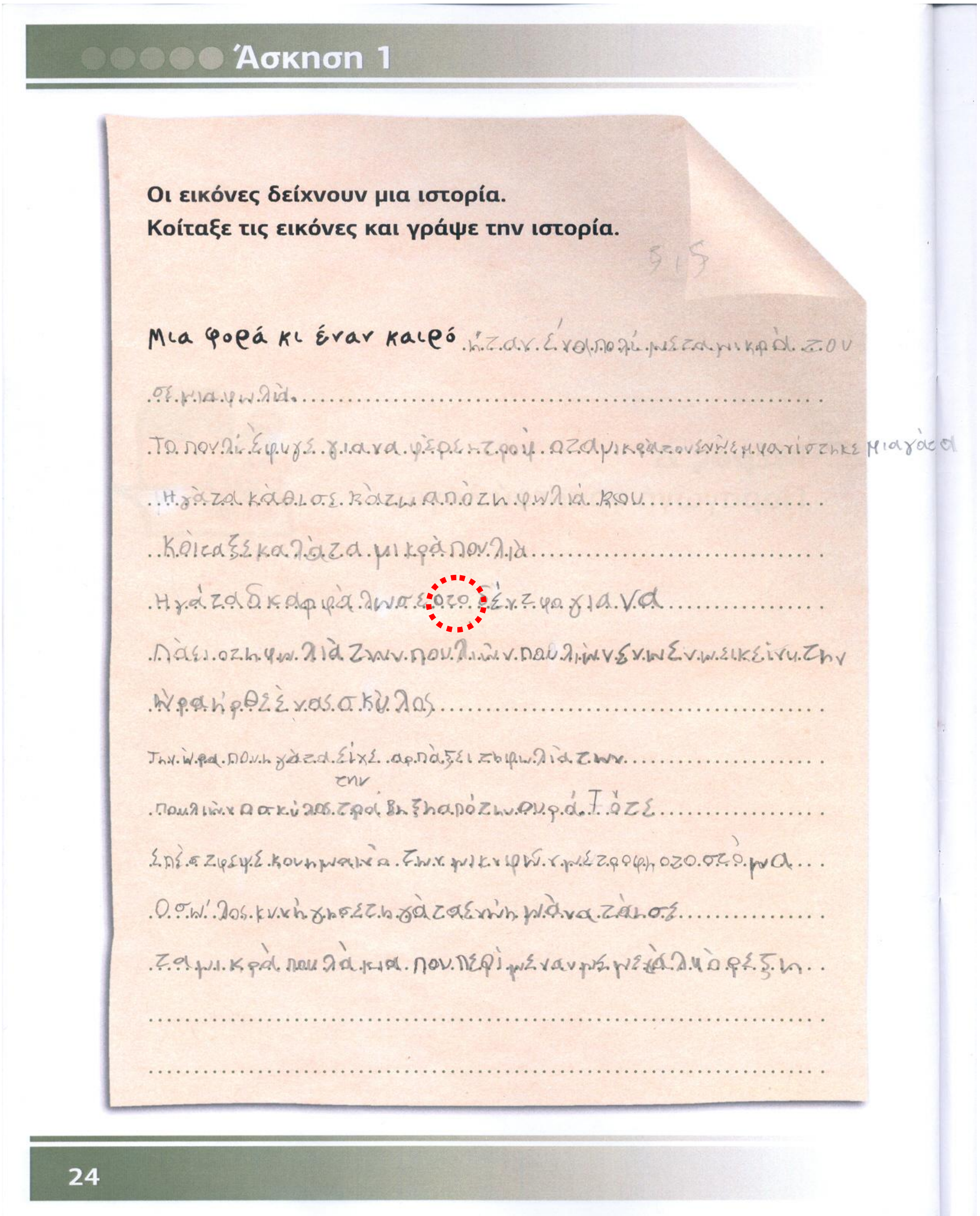
.....

.....

.....

21

Γράφημα 1. Κείμενο 1



Άσκηση 1

Οι εικόνες δείχνουν μια ιστορία.
Κοίταξε τις εικόνες και γράψε την ιστορία.

Μια φορά κι έναν καιρό ήταν ένα πουλί με τα μικρά του σε μια φωλιά.

Το πουλί έφυγε για να φέρει τροφή [illegible] εμφανίστηκε μια γάτα

Η γάτα κάθισε κάτω από τη φωλιά [illegible]

Κοίταξε [illegible] μικρά πουλιά

Η γάτα [illegible] για να

[illegible] την

ώρα ήρθε ένας σκύλος

Την ώρα που η γάτα είχε αρπάξει [illegible]

την

[illegible] Τότε

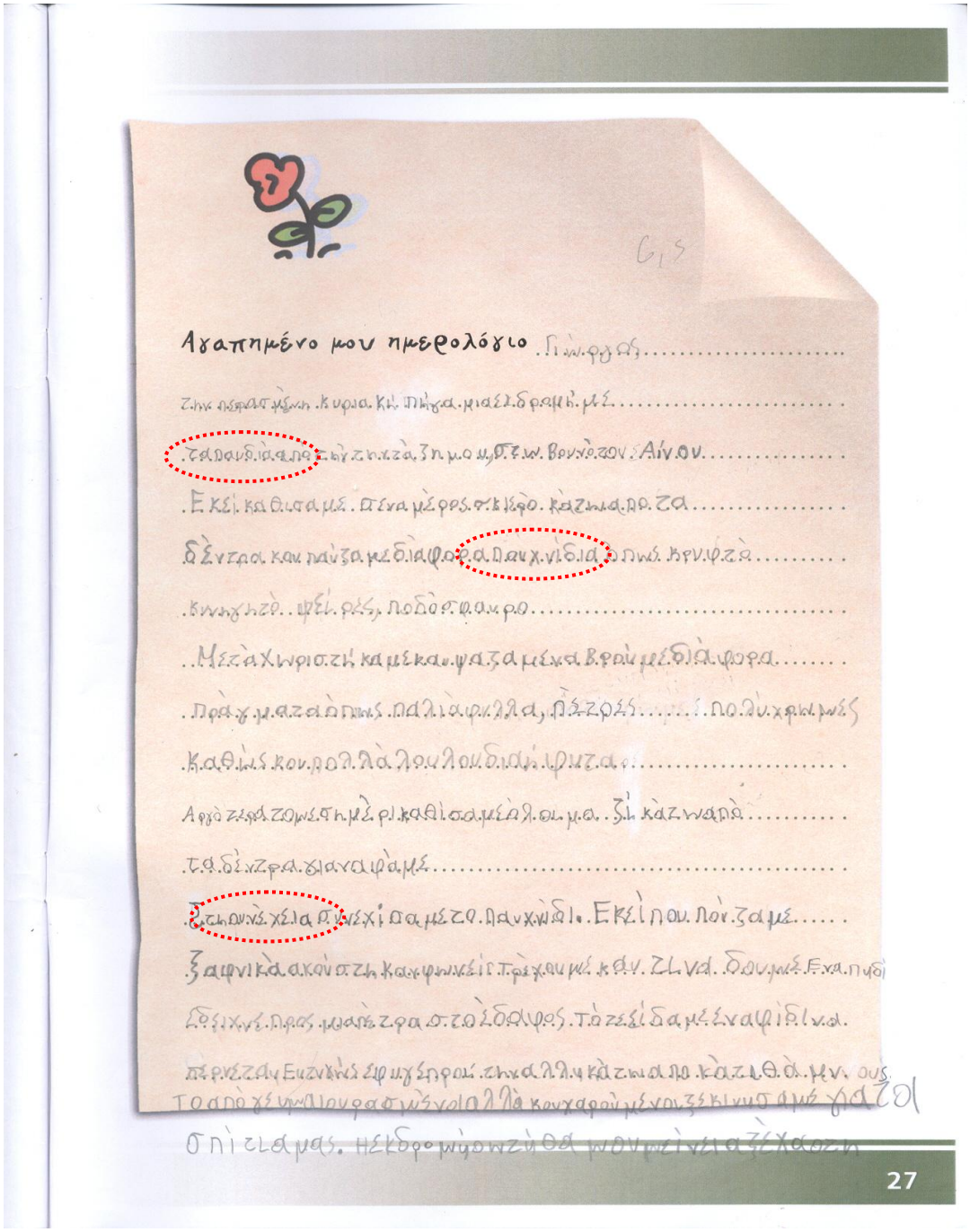
[illegible]

[illegible]

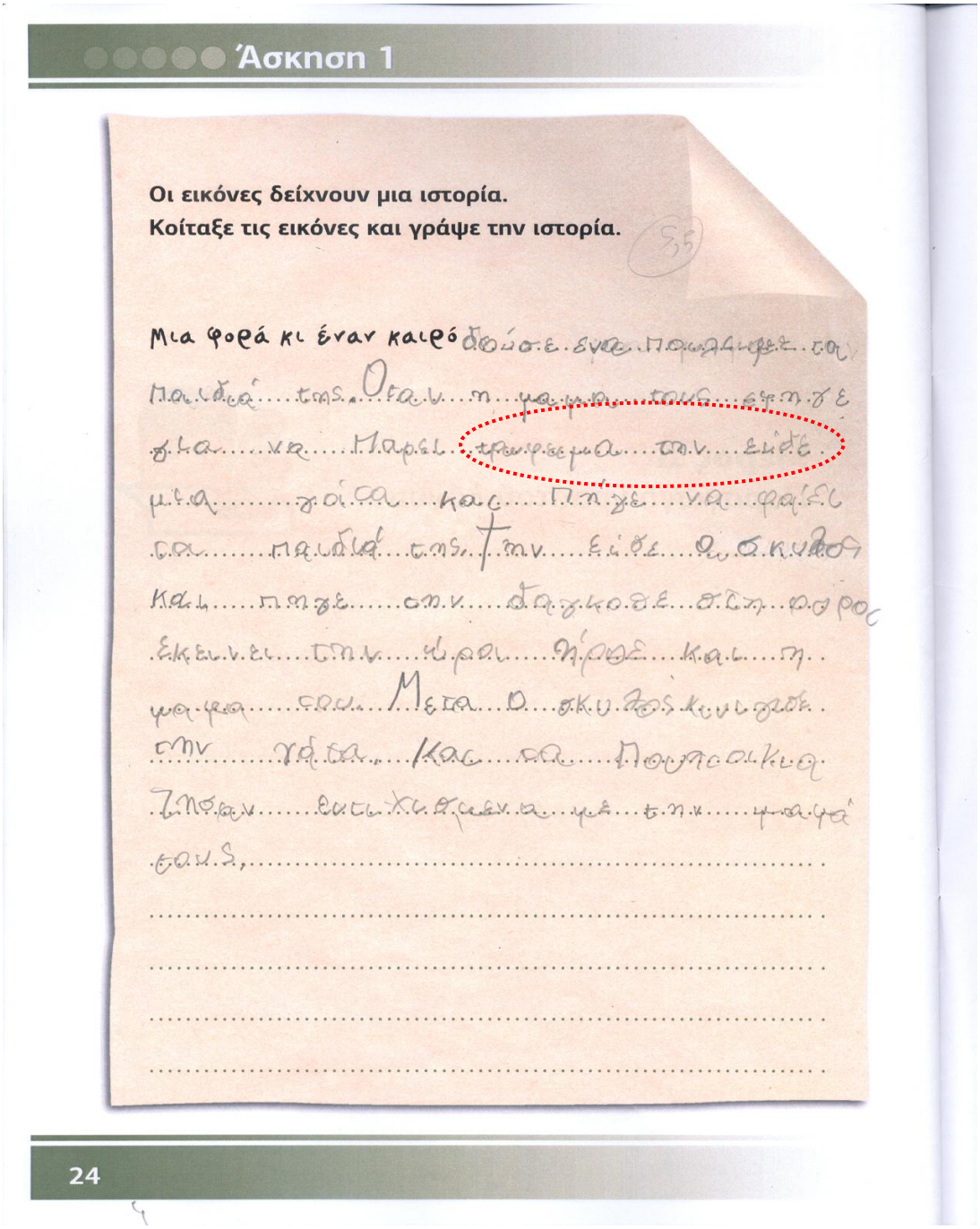
[illegible]

24

Γράφημα 2. Κείμενο 2



Γράφημα 3. Κείμενο 3



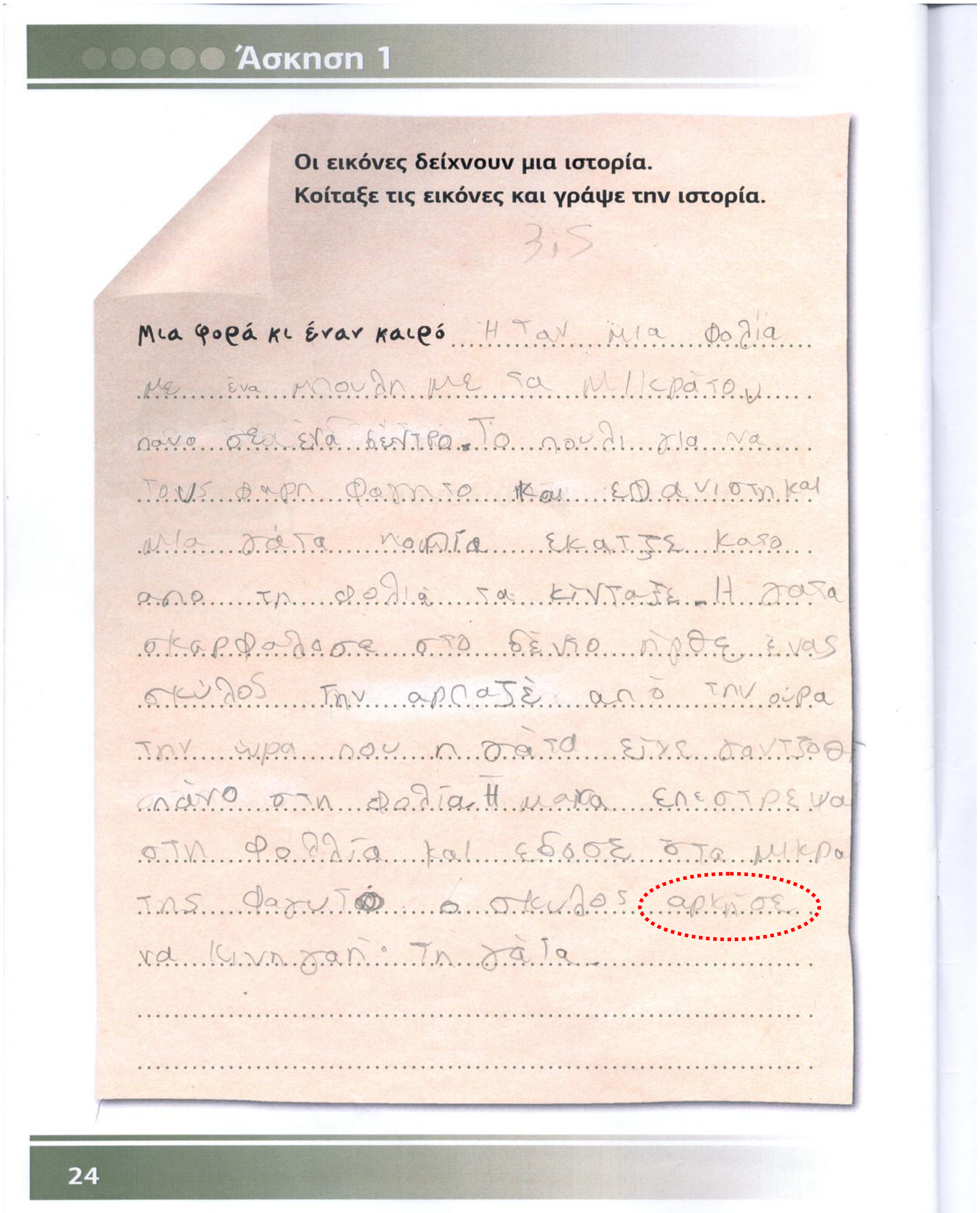
Άσκηση 1

Οι εικόνες δείχνουν μια ιστορία.
Κοίταξε τις εικόνες και γράψε την ιστορία.

Μια φορά κι έναν καιρό [illegible] Όταν η μαμά τους έφτιαξε
για να πάρει τρόφιμα αν είδε
μια γάτα και τους πήγε να φάει
τα παιδιά της την είδε φώναξαν
και πήγε αν δαγκώσει στη φορά
εκείνη την ώρα ήρθε και η
μαμά τους Μετα ο σκύλος κυνηγούσε
την γάτα και τα πουλάκια
ζήσαν ευτυχισμένα μέχρι τα γηρατειά
τους.

24

Γράφημα 4. Κείμενο 4


Άσκηση 1

Οι εικόνες δείχνουν μια ιστορία.
Κοίταξε τις εικόνες και γράψε την ιστορία.

3,5

Μια φορά κι έναν καιρό Ηταν μια φολία
με ενα πουλη με τα μικράτου
πανο στα ενα δεντρο. Το πουλι για να
τους φερη φαγητο και εφανιστηκαι
μια γατα πουλια εκατσε κατω
απο τη φολιά να κινταξε. Η γατα
σκαρφαλασε στο δενρο ηρθε ενας
σκύλος την αρπαξε απο την ούρα
την ώρα που η γατα ειχε φαντσοθ
πανω στη φολια. Η μαμα επεστρεψα
στη φολλια και εδοσε στα μικρα
της φαγυτο ο σκυλος αρκησε
να κυνηγαη τη γατα.

24

Γράφημα 5. Κείμενο 5

3.2 Πλαίσιο επισημείωσης του ΣΚΕΙΜΑΘ

Ένα σημαντικό κομμάτι της πορείας του ΣΚΕΙΜΑΘ είναι η συμφωνία για ένα πλαίσιο επισημείωσης λαθών το οποίο θα αποτελέσει τη βάση για την ανάλυση των λαθών στα τελευταία στάδια του προγράμματος. Ο κύκλος δημιουργίας του ΣΚΕΙΜΑΘ και ανάλυσης λαθών που καθοδηγεί τα βήματα της υποδράσης 1.5 βασίζεται σε κάποιες κοινές αρχές που ακολουθούν διεθνή προγράμματα με τον ίδιο σκοπό. Η πορεία του προγράμματός μας υιοθετεί μεθοδολογικά

τα πρότυπα του προγράμματος Free Text για το ΣΚΜ FRIDA στη Γαλλία το 1998, το οποίο ακόμη αποτελεί σημείο αναφοράς για τα ΣΚΜ. Πιο συγκεκριμένα, ακολουθώντας τις κύριες μεθοδολογικές αρχές του Free Text, η υποδράση 1.5 ακολουθεί τα εξής βήματα για τη δημιουργία και ανάλυση του ΣΚΕΙΜΑΘ:

- χειρωνακτική εύρεση των λαθών στο ΣΚΕΙΜΑΘ,
- επεξεργασία και συμφωνία για το σύνολο ετικετών που θα απαρτίζουν το πλαίσιο επισημείωσης λαθών,
- εισαγωγή ετικετών λαθών και διορθώσεων στα αρχεία κειμένων,
- ανάκτηση/εξαγωγή λιστών συγκεκριμένων τύπων λαθών και στατιστική ανάλυση λαθών, και
- γλωσσική ανάλυση των σημαντικότερων τύπων λαθών, βάσει συμφραστικών πινάκων (concordance-based)/ συνταυτίσεων.

Το πλαίσιο επισημείωσης των λαθών του ΣΚΕΙΜΑΘ σχεδιάστηκε λαμβάνοντας υπόψη από τη μια πλευρά τη διαγλώσσα των μαθητών κι από την άλλη τις ανάγκες των χρηστών. Ακολουθώντας την Granger (2003), στοχεύσαμε σε ένα πλαίσιο επισημείωσης το οποίο είναι:

I. Διαφωτιστικό (informative) και ταυτόχρονα διαχειρίσιμο (manageable): να μπορεί να είναι τόσο αναλυτικό ώστε να παρέχει χρήσιμες πληροφορίες για τα λάθη των μαθητών, αλλά όχι τόσο αναλυτικό που να μην μπορεί να το διαχειριστεί ο επισημειωτής.

II. Επαναχρησιμοποιήσιμο (reusable): οι κατηγορίες θα πρέπει να είναι αρκετά γενικές, ώστε να μπορούν να χρησιμοποιηθούν για διαφορετικές γλώσσες.

III. Ευέλικτο (flexible): θα πρέπει να επιτρέπει άμεση πρόσβαση για αλλαγές (προσθήκη/ αφαίρεση ετικετών) πάνω στα επισημειωμένα κείμενα.

IV. Συνεπές (consistent): να μην υπάρχουν αντιφάσεις στις επισημειώσεις των κειμένων, όταν αναμειγνύονται περισσότεροι του ενός επισημειωτές. Επομένως, απαιτείται ένα εγχειρίδιο χρήσης με λεπτομερή περιγραφή των αρχών επισημείωσης και των κατηγοριών λαθών, με την παραδειγματική εφαρμογή τους.

Σε σχέση με την κατηγοριοποίηση των λαθών, το ΣΚΕΙΜΑΘ τα διαχωρίζει πρακτικά σε δύο κύριες κατηγορίες ακολουθώντας τους Dulay, Burt and Krashen (1982):

- Λάθη βασισμένα σε γλωσσολογικές κατηγορίες (μορφολογίας, σύνταξης κτλ.)
- Τρόπος πραγματοποίησης του λάθους (παράλειψη, πρόσθεση, παραμόρφωση δεδομένων κτλ.).

Επιπλέον, η κατηγοριοποίηση γίνεται ιεραρχικά και περιλαμβάνει τρία στρώματα ιεράρχησης κατά τρόπο παρόμοιο με το ΣΚΜ FRIDA του προγράμματος Free Text:

- Τομέας λαθών (error domain)
- Κατηγορία λαθών (error category)
- Γραμματική κατηγορία της λέξης (word category)[2]

Στον πίνακα 2 περιγράφεται λεπτομερώς το πλαίσιο επισημείωσης λαθών της Υποδράσης 1.5 μετά από τη συνεργασία των μελών της ομάδας.

Τομέας λαθών		**Κατηγορία Λαθών**		**Παραδείγματα**
Σύμβαση / Κώδικας	***Επεξήγηση***	***Σύμβαση / Κώδικας***	***Επεξήγηση***	
_ΟΡΘ	Ορθογραφία (Υπάρχει σωστή φωνολογική αναπαράσταση- εντοπίζονται μόνο τα ορθογραφικά λάθη)	**-ΛΕΞ**	Λάθη στο θέμα/λεξικά μορφήματα	φολιά/ ευδομάδα/ φαγιτο/ κε ανεβικιε/ πέζαμε/ προί
		-ΓΡΑΜ	Λάθη στα κλιτικά επιθήματα/γραμματικά μορφήματα	παρη/ δοσι/ ταηση/ στω εκδρομι/ κιταξι
		-ΠΑΡ	Λάθη στα παραγωγικά μορφήματα	Σχοληο
		-ΣΥΓΧ	Λάθη συγχώνευσης (agglutination) Στις περιπτώσεις της συγχώνευσης εντοπίζουμε **μόνο** την συγχώνευση και δεν επισημειώνουμε άλλα λάθη	τιλες/ παραπολι/ τινα/ μαδεν/ σαφινο μετα/ τιλες
		-ΚΕΦ	Χρήση κεφαλαίων αντί πεζών	...ήταν Δύο πουλάκια.../ ...έναν καιρό Ήταν μια.../... με φωνάζει Η μάνα.../ ...στην θαλασσά Βοβειθίκα. .. /την τραβιξε απο την ουρα της Εκεινι την ωρα…. τα Φαγιτο/
		-ΠΕΖ	Χρήση πεζών αντί για κεφαλαία	στην Βόρια αμερίκη/ σο μεσολογγι/ γιώργο/ αλβανια/ τότε επέστφεψε και η μαινό των μικνφών με τροφη οτο στόμα. ο σκυλος κυνήγησε τη γάτα ενω η μάνα…..

[2] Στην επόμενη φάση του προγράμματος θα αποφασιστεί ποιο σύνολο ετικετών είναι πιο χρήσιμο σε σχέση με τη λεπτομέρεια περιγραφής των γραμματικών κατηγοριών.

		-ΚΩΔ	Χρήση λατινικού αντί για ελληνικό αλφάβητο	
_ΓΡΑΦ	Γραφηματικά λάθη (Λάθη στη μορφή της λέξης, δεν υπάρχει σωστή φωνολογική αναπαράσταση	**-ΛΕΞ**	Λάθη λεξικών μορφημάτων	πηλί (αντί για πουλί)/ δεντο (αντί για δέντρο) / εγρομι (αντί για εκδρομή)/ πολί (αντί για πουλί)/ μοχθρις (αντί για μοχθηρής)
		-ΓΡΑΜ	Λάθη γραμματικών μορφημάτων	έκατζε/ ...ήθελε να τους φα.../ πηγαμ
_ΤΟΝ	Λάθη στον τονισμό	**- ΠΡΟΣ**	Πλεονασμός τόνου (Oversupply)	βρεί/ ναί/ τήν/ τρείς χθές/ πίγάμε
		-ΠΑΡ	Παράλειψη τόνου (Omission)	σκυλος/ φαει/ δαγοσε/ εκει ξεκινησαμε/ένα
		-ΑΝΤΙ	Αντικατάσταση (Confusion/substitution)	επέσα/ γονόις/ ερχέτε αύτοι/ ούρα
_ΣΥΜΦ	Λάθη στη συμφωνία	**-ΑΡΙ**	Λάθη στον Αριθμό	...ηταν ωραια το φαγιτο.../ στο μορα/ η λιμνες/ αυτοί η εγδρόμη τα Φαγιτο
		-ΓΕΝ	Λάθη στο Γένος	η πουλί/ στους πουλιά/ το ουρά/ ενας περιστέρι
		-ΠΡΟΣ	Λάθη στο πρόσωπο, ανάλογα με την γραμματική κατηγορία των όρων που συμφωνούν	τον (αντί για «με»)/ εγώ τα περνα/ η μανα επεστρεψα
		-ΠΤΩ	Λάθη στην Πτώση	ενας σκιλο
_ΓΕΝ	Λάθη στην απόδοση γένους			μια δέντρα/ στον δερντο/ τα παιδια της (αντί για του)/ ένα ντήσκο
_ΟΨΗ	Λάθη που αφορούν την όψη του ρήματος	**-ΣΥΝ**	Όταν χρησιμοποιείται συνοπτικό στη θέση μη συνοπτικού	...άρχισαμε να παίξουμε.../ ...τελειώσαμε να παίξουμε...
		-ΜΣΥΝ	Όταν χρησιμοποιείται μη συνοπτικό στη θέση συνοπτικού	...Εκεί πήγαμε και πέζαμε.../ ...ένας σκύλος την αρπαζέ.../ ...ηταν πηλί ορεα ηχε δεντρα λούλουδια διαβαζαμε παραμιθια.../ ...Επεζαμε με τους φηλουσ μου γελάσαμε.../

						και εγω εχω πολα μαθήματα και να διαβαζω
_ΧΡ	Λάθη στον χρόνο του ρήματος	**-ΠΡΛΘ**	όταν χρησιμοποιείται παρελθοντικός αντί μη παρελθοντικού χρόνου			
		-ΜΠΡΛΘ	όταν χρησιμοποιείται μη παρελθοντικός αντί παρελθοντικού χρόνου			...εχώ Πάη (αντί για πήγα).../ ...όταν πιγα εκδρομι ηταν σουπερ μου αρεσει (αντί για άρεσε).../ ...Η λιμνες, στιν εκδρομι δεν μου αρεσουν (αντί για άρεσαν) τα μαγαζια δεν μου αρεσουν γιατι ητανε...
_ΛΕΞΗ	Λάθη προσθήκης ή παράλειψης λέξεων περιεχομένου (ουσιαστικά, επίθετα, ρήματα)	**-ΠΑΡ**	Παράλειψη Λέξης			
		-ΠΡΟΣ	Προσθήκη Λέξης			...πέζαμε μετά αγόρια κινηγητό (*πέζαμε) ποδοσφερο βολεη
_ΣΤΙΞ	Λάθη στα σημεία στίξης	**-ΠΑΡ**	Παράλειψη Στίξης			...πέζαμε μετά αγόρια κινηγητό ποδοσφερο βολεη... την τραβιξε απο την ουρα της Εκεινι την ωρα…./ καθίσαμε όλοι μαζί κάτω από τα δέντρα για να φάμε Στη συνέχεια συνεχίσαμε το παιχνίδι
		ΠΡΟΣ	Προσθήκη Στίξης			ειδαμαι ολιτιν,ακροπολι/ ο μπαμπάς μου
		-ΑΝΤΙ	Αντικατάσταση στίξης			...κατο απο την φωλλια, Μετα... (αντί τελείας)
_ΚΕΙΜ	Επισημείωση κειμενικών δεικτών συνεκτικότητας κειμένου					...έρχετε η γάτα και και φέυχι.../ ... Και πιο μετα εινταν πανο στο δενρο...
_ΣΟΡ	Λάθη στην σειρά των όρων					...Το πουλί για να τους φωρη φαγητο και εμφανιστηκαι
_ΑΡΘ	Λάθη που αφορούν το	**_ΟΡΙ**	Οριστικό	**-ΠΑΡ**	-Παράλειψη	και («η») γατα εινταν/ ...(«Το»)

	άρθρο		άρθρο			Βράδυ 7 («η») ώρα πήγα...
				-ΠΡΟΣ	-Προσθήκη	όταν πειγαμε ***στι*** εγρομι
				-ΑΝΤΙ	Αντικατάσταση	
		_ΑΟΡ	Αόριστο άρθρο	**-ΠΑΡ**	-Παράλειψη	Μια δέντρα πουλί και τα πεδαιά
				-ΠΡΟΣ	-Προσθήκη	
				-ΑΝΤΙ	Αντικατάσταση	
_Ν	Τελικό –ν (υποχρεωτικά μόνο πριν από φωνήεν)					στο(ν) Άγιο Πέτρο/ στη ουρα
_ΚΛΙΤ	Λάθη προσθήκης ή παράλειψης κλιτικών	**-ΠΑΡ**	Παράλειψη κλιτκού			...μαζι μετο μικρα («του»)...
		-ΠΡΟΣ	Προσθήκη κλιτικού			
		-ΑΝΤΙ	Αντικατάσταση κλιτκού (ως προς την πτώση του κλιτικού)			
_ΠΡΟΒΛ	Για περιπτώσεις αβέβαιης ταξινόμησης ως λαθών					Ρέριξε, Κια

Πίνακας 2. Πλαίσιο επισημείωσης λαθών του ΣΚΕΙΜΑΘ

3.3 Πρακτικές επισημειώσεων

Το τεχνικό κομμάτι της επισημείωσης είναι το ίδιο κρίσιμη διαδικασία όσο και το πλαίσιο επισημείωσης λαθών. Αν και στα πρώτα ΣΚΜ η επιλογή της γλώσσας κωδικοποίησης των λαθών είχε ως κύριο κριτήριο το κατά πόσο ακολουθούσε τις διεθνείς τάσεις προτυποποίησης των ΣΚ, σήμερα οι τεχνικές δυνατότητες των γλωσσών σήμανσης συνδέονται άμεσα με:

- τη δυνατότητα χειρισμού,
- ανάλυσης και
- χρήσης των επισημειωμένων δεδομένων.

Το 1998 το World Wide Web Consortium, η κύρια κοινοπραξία για την επικύρωση προτύπων στο διαδίκτυο, παρουσιάζει τη γλώσσα σήμανσης XML, ένα υποσύνολο της δύσχρηστης γλώσσας Standard Generalized Markup Language (SGML) που είχε δημιουργηθεί με

πολύ διαφορετική στόχευση από ό,τι η XML. Έκτοτε, μια σειρά προτύπων αναπτύχθηκαν με βάση την XML.[3]

Η έκρηξη στην ανταλλαγή πληροφοριών μέσα στο διαδίκτυο και η σήμανσή τους μέσα από την XML αναγνωρίστηκαν άμεσα και στην κοινότητα της γλωσσολογίας σωμάτων κειμένων και η μεγάλη πλειονότητα των ΣΚ που παρουσιάζονται τα τελευταία χρόνια χρησιμοποιούν την XML ως τη γλώσσα κωδικοποίησης των δεδομένων.

Το ΣΚΕΙΜΑΘ επισημειώνεται σε XML και κύριο μέλημά του σε σχέση με την ερμηνεία και περιγραφή των λαθών των μαθητών είναι:

- **να αποτυπώσει την εικόνα των εξωτερικών χαρακτηριστικών της διαγλώσσας των μαθητών, χωρίς να μπει σε διαδικασία ερμηνείας.**

Αυτή η κεντρική επιλογή του ΣΚΕΙΜΑΘ το διαφοροποιεί από την προηγούμενη προσπάθεια του Τζιμώκα (2010) για τη δημιουργία ελληνικού ΣΚΜ. Το ΣΚΜ του Τζιμώκα (2010) αποτελείται από περίπου 65.000 λέξεις και 291 κείμενα και είναι η πρώτη μεγάλη συστηματική προσπάθεια κωδικοποίησης των λαθών μαθητών με την ελληνική ως ξένη/δεύτερη γλώσσα. Το πλαίσιο επισημείωσης λαθών του Τζιμώκα (2010) είναι πολύ λεπτομερειακό και βασίζεται σε λογισμικό επισημείωσης και επικύρωσης το οποίο περιορίζει τις δυνατότητες κωδικοποίησης πολλαπλών φαινομένων και ταυτόχρονα κατευθύνει την ίδια την κατηγοριοποίηση του πλαισίου επισημείωσης λαθών. Συγκεκριμένα, η επισημείωση γίνεται πάνω στα ψηφιοποιημένα κείμενα με τη λεγόμενη in-line επισημείωση με δύο κύρια μειονεκτήματα:

- Η σύνταξη της XML δεν επιτρέπει διασταυρωμένες επισημειώσεις, δηλαδή περιπτώσεις επισημείωσης όπου τα όρια της μιας επισημειώσης περιλαμβάνονται μόνο εν μέρει σε τμήμα άλλης επισημείωσης (λ.χ. <ετικέτα1>....<ετικέτα2>....</ετικέτα1>...</ετικέτα2>).

[3] Ενδεικτικά:
DOM Level 1 V1.0 (Οκτώβριος 1998)
XML Namespaces V1.0 (Ιανουάριος 1999)
XPath V1.0 (Νοέμβριος 1999)
XSLT V1.0 (Νοέμβριος 1999)
XHTML V1.0 (Ιανουάριος 2000)
XML Schema V1.0 (Μάιος 2001)
XLink V1.0 (Ιούνιος 2001)
XPointer V1.0 (Σεπτέμβριος 2001)
XSL V1.0 (Οκτώβριος 2001)
XML Information Set V1.0 (Οκτώβριος 2001)
XPath 2.0 WD (Απρίλιος 2002)...

- Απαιτούνται δύο επισημειώσεις σε περιπτώσεις πολυλεξικών μονάδων αν υπάρχει ανάγκη για αναφορά σε μια από τις λέξεις (λ.χ. “παραδείγματος χαρα” αντί του “παραδείγματος χάρη” για αναφορά μόνο στη δεύτερη λέξη της παγιωμένης φράσης).

Επιπλέον, το ΣΚΜ του Τζιμώκα (2010) δεν ακολουθεί τις σύγχρονες βελτιστοποιημένες διεθνείς προδιαγραφές επισημειωμένων ΣΚ με την τακτική in-line, όπως την σύμβαση TEI (Text Encoding Initiative) Guidelines (π.χ. το ΣΚ jos100k από τους Erjavec Tomas κ.α. 2010). Έτσι, αν και η εμβέλεια των δεδομένων και η συστηματικότητα του ΣΚΜ του Τζιμώκα (2010) είναι εντυπωσιακή, ο τεχνικός σχεδιασμός όσον αφορά την επισημείωση επηρέασε

1. την ευελιξία του πλαισίου επισημείωσης, καθώς περιέχει ένα μεγάλο αριθμό κατηγοριών με μεγάλη περιγραφική λεπτομέρεια εξαιτίας της δυσκολίας κάλυψης φαινομένων με τα διαθέσιμα εργαλεία του,
2. την εμβέλεια χρήσης και ανάλυσης των δεδομένων και
3. την συμβατότητα με άλλες σύγχρονες πλατφόρμες και μορφές επισημείωσης επιτρέποντας παραλληλισμό με άλλα ΣΚΜ από άλλες γλώσσες.

Το ΣΚΕΙΜΑΘ βασίζεται στη δεύτερη πρόσφατη και συνεχώς αύξουσα τάση για επισημείωση των δεδομένων έξω από το ψηφιοποιημένο αρχείο του κειμένου, τη λεγόμενη επισημείωση standoff. Η κύρια αρχή αυτής της πρακτικής επισημείωσης απαντά στα κλασικά προβλήματα που αναφέρθηκαν πιο πάνω και διαχωρίζει το αυθεντικό ψηφιοποιημένο κείμενο και τις επισημειώσεις, οι οποίες τώρα κάνουν αναφορά στο σημείο του κειμένου που στοχεύουμε να επισημειωθεί.

Στα πλαίσια προτυποποίησης της πρακτικής standoff, η ομάδα προτυποποίησης ISO/TC 37/SC4 (cf.http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37326) δημιούργησε το πρότυπο του Linguistic Annotation Framework (LAF). Το LAF είναι μια προκαθορισμένο και προσυμφωνημένο υποσύνολο της XML, το οποίο έχει σχεδιαστεί ειδικά για την γλωσσολογική κωδικοποιήση δεδομένων.

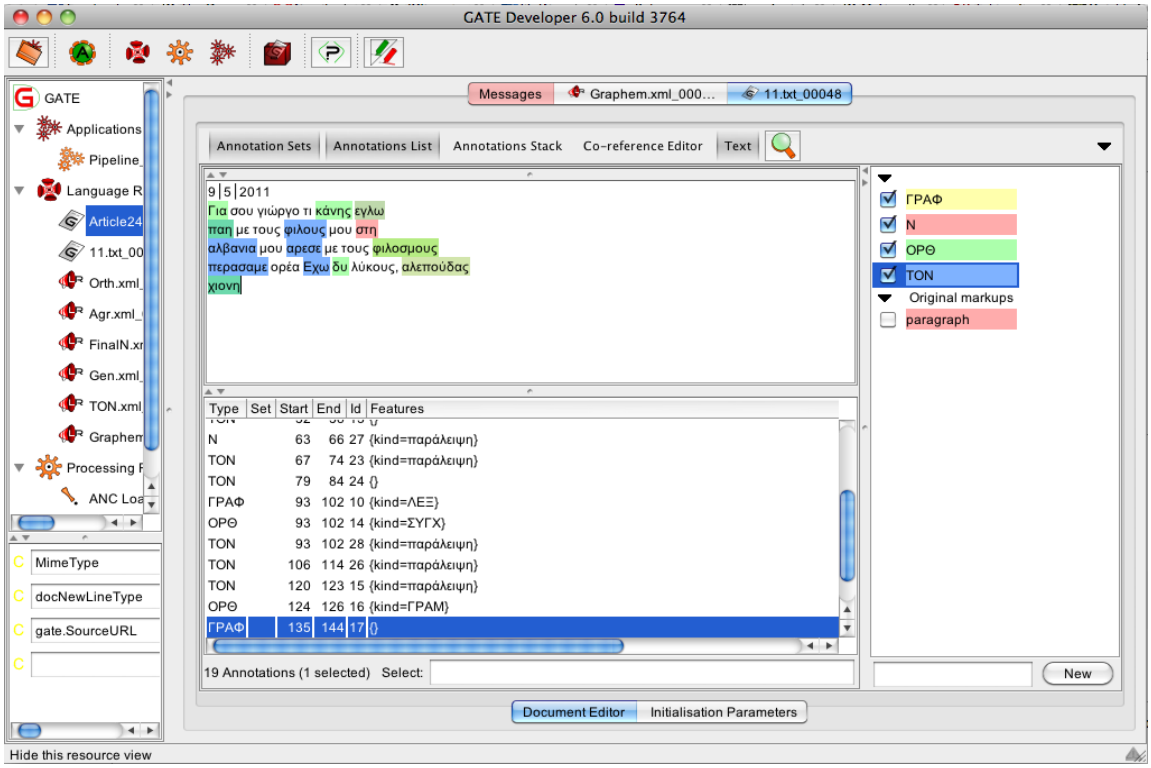
Τις τελευταίες δεκαετίες, είχε καθιερωθεί πληθώρα γλωσσών σήμανσης και μορφοποιήσεων για τα ΣΚ και η διεθνής κοινότητα αναγνώρισε σχετικά πρόσφατα την ανάγκη για τη δημιουργία μιας κοινής γλώσσας γλωσσολογικών επισημειώσεων. Το LAF δίνει το πλαίσιο γύρω από το οποίο οι ερευνητές της γλώσσας μπορούν να προσθέσουν τις δικές τους ετικέτες, προκειμένου να συνεισφέρουν όλοι στον παραλληλισμό και την ευθυγράμμιση των επισημειωμένων ΣΚ, επιτρέποντας έτσι την συγκριτική ανάλυση δεδομένων σε διάφορα επίπεδα ανάλυσης και από πολλές διαφορετικές γλώσσες. Υιοθετώντας το LAF, το ΣΚΕΙΜΑΘ γίνεται μέλος μιας συνεχώς αυξανόμενης διεθνούς κοινότητας ΣΚ, στην οποία ανήκει ανάμεσα σε άλλα, το American National Corpus (ANC). Η μορφοποίηση του LAF υλοποιείται τεχνικά μέσα από το Graph Annotation Framework (GraF) (Ide and Romary 2007). Το GraF είναι η μορφοποίηση για

την κωδικοποίηση γλωσσολογικών κατηγοριών ως ετικέτες σε τυπικές δομές χαρακτηριστικών (typed feature structures) συνδεδεμένες με τα αυθεντικά αρχεία των παραγωγών.

3.4 Επισημείωση και χρήση του ΣΚΕΙΜΑΘ μέσα στο περιβάλλον GATE

Το περιβάλλον GATE είναι μια πλατφόρμα δημιουργίας και αξιοποίησης εφαρμογών επεξεργασίας φυσικών γλωσσών. Το GATE υποστηρίζει το πλαίσιο επισημείωσης λαθών του ΣΚΕΙΜΑΘ και παρέχει λειτουργικότητα στην επισημείωση με τα χαρακτηριστικά χρωματισμού και εύκολης διόρθωσης και συγγραφής πάνω στο κείμενο. Επιπλέον, συνδυάζει το πλεονέκτημα του διαχωρισμού του αρχικού αρχείου των παραγωγών από το αρχείο τεμαχισμού και τα αρχεία αναφοράς και επισημείωσης, καθώς οπτικά συνδυάζει τα τρία είδη αρχείων. Έτσι, ο χρήστης θα μπορεί αναπόσπαστα να ασχολείται με το αντικείμενο έρευνας και όχι με εργαλεία συνδυασμού των αρχείων, προκειμένου να αποκτήσει δυνατότητα εποπτείας όλων των δεδομένων.

Στο δεξιό μέρος του πίνακα 3 διακρίνονται οι κατηγορίες λαθών που έχουν “φορτωθεί” στο σύστημα και είναι διαθέσιμες για επισημείωση. Στο κεντρικό κομμάτι της οθόνης διακρίνονται στο πάνω τμήμα το κείμενο με τα επισημειωμένα λάθη σε διαφορετικά χρώματα που αντιστοιχούν με το χρώμα στα δεξιά. Επιπλέον, στο κάτω τμήμα φαίνεται και η υποκατηγορία του λάθους μαζί με την εμβέλειά του σε χαρακτήρες. Για παράδειγμα, στον πίνακα 3 με μπλε χρώμα έχουν χρωματιστεί τα λάθη τονισμού, ενώ όταν ο χρήστης περιηγηθεί σε κάποιο από τα λάθη, τότε στο κάτω μέρος μαρκάρεται η εμβέλεια (αριθμημένων χαρακτήρων) και η υποκατηγορία με την οποία είναι επισημειωμένο. Στην αριστερή πλευρά, εμφανίζεται κομμάτι της λειτουργικότητας του GATE με τα ονόματα των αρχείων τα οποία κωδικοποιούν το πλαίσιο επισημείωσης λαθών και “εμπλέκονται” στην επαλήθευση των επισημειώσεων ως ένα είδος γραμματικής λαθών. Το GATE παρέχει μια αξιόπιστη εικόνα του χάρτη λαθών του μαθητή. Στην επόμενη φάση του προγράμματος η υποδράση 1.5 θα χρησιμοποιήσει τα εργαλεία java του ΣΚ ANC σε όλο το εύρος τους για την επισημείωση με την πρακτική standoff. Το σχεδιαζόμενο ΣΚΕΙΜΑΘ ακολουθώντας την πρακτική επισημείωσης standoff παρέχει τη δυνατότητα σε μελλοντικούς ερευνητές να χρησιμοποιήσουν τα δεδομένα χωρίς να είναι αναγκασμένοι να ασχοληθούν με τις αρχικές επισημειώσεις ΣΚΕΙΜΑΘ, καθώς αυτές θα βρίσκονται σε ξεχωριστό αρχείο.



Πίνακας 3. Παράδειγμα επισημειωμένου κειμένου στο περιβάλλον GATE με το πλαίσιο επισημείωσης λαθών του ΣΚΕΙΜΑΘ

4 Αναφορές

Ξενόγλωσσες

Aarts, J. & Granger, S. 1998. "Tag sequences in learner corpora: A key to interlanguage grammar and discourse". In S. Granger (ed.), *Learner English on computer*. London: Addison Wesley Longman, pp. 132-141.

Abbott, G. 1980. "Towards a more rigorous analysis of foreign language errors". *IRAL* 18 (2): 121-134. Cambridge University Press. 2006. *Cambridge Learner Corpus*. [Available at http://www.cambridge.org/elt/corpus/learner_corpus.htm]

Altenberg, B. 2002. "Using bilingual corpus evidence in learner corpus research". In S. Granger, J. Hung & S. Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, pp. 37-54.

Atkins, S., Clear, J. & Ostler, N. 1992. "Corpus design criteria". *Literary and Linguistic Computing*, 7: 1-16.

Biber, D. 1993. "Representativeness in corpus design". *Literary and Linguistic Computing*, 8 (4): 243-257.

Cobb, T. 2003. "Analyzing late interlanguage with learner corpora: Québec replications of three European studies". *The Canadian Modern Language Review / La Revue canadienne des langues vivantes*, 59 (3): 393-423.

Corder, S. P. 1974. "Error analysis". *The Edinburgh Course in Applied Linguistics (Vol.3).* Eds. J. Allen & S. P. Corder. London: Oxford University Press, 122-154.

Cowan, R., Choi, H. E. & Kim, D. H. 2003. "Four questions for error diagnosis and correction in CALL". *CALICO Journal*, 20 (3): 451-463.

Dagneaux, E. et al. 1996. *Error Tagging Manual Version 1.1.* Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Dagneaux, E. et al. 1998. "Computer-aided error analysis". *System* 26: 163-174.

De Cock, S. 1998. "A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English". *International Journal of Corpus Linguistics, 3: 59-80.*

de Haan, P. 1984). "Problem-oriented tagging of English corpus data". In J. Aarts & W. Meijs (eds), *Corpus linguistics: Recent developments in the use of computer corpora*, London: Addison Wesley Longman, pp. 123-139.

de Haan, P. 2000. “Tagging non-native English with the TOSCA-ICLE tagger”. In C. Mair & M. Hundt (eds), *Corpus linguistics and linguistic theory*, Amsterdam: Rodopi, pp. 69-79.

de Mönnink, I. 2000. “Parsing a learner corpus”. In C. Mair & M. Hundt (eds), *Corpus linguistics and linguistic theory*, Amsterdam: Rodopi, pp. 81-90.

Díaz-Negrillo, A. & García-Cumbreras, M. A.. Forthcoming. “A tagging tool for error analysis on learner corpora”. *ICAME Journal*.

Díez Prados, M. et al. 2006. “The ICLE error tagging project: analysis of Spanish EFL writers”. Paper presented at the *Fourth International Contrastive Linguistics Conference*, Santiago de Compostela, Spain, 19-23 September.

Ellis, R. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, R. & Barkhuizen, G. 2005. *Analysing Learner Language*. Oxford University Press, Oxford.

Fitzpatrick, E. & Seegmiller, M. S. 2004. “The Montclair electronic language database project”. In U. Connor & T. A. Upton, eds., *Applied Corpus Linguistics. A Multidimensional Perspective,* pp. 223-237. Amsterdam: Rodopi.

Flowerdew, L. & Tong, A. K. K. (eds.) 1994. *Entering text*. Hong Kong: Language Centre, Hong Kong University of Science and Technology, and Department of English, Guangzhou Institute of Foreign Languages, pp.157-165.

Garside, R., Leech, G. & McEnery, A. (eds.) 1997. *Corpus annotation: Linguistic information from computer text corpora.* London: Longman.

Gass, S. M. & Selinker, L. 2001. *Second language acquisition: An introductory course*. Mahwah, NJ: Lawrence Erlbaum.

Granger, S. & Tyson, S. 1996. “Connector usage in the English essay writing of native and non-native EFL speakers of English”. *World Englishes*, 15: 19-29.

Granger, S. 1998. “Computer-aided error analysis”. *System*, 26:163–174.

Granger, S. 1998. “*Learner English on computer*”. London: Addison Wesley Longman.

Granger, S. 1998. “The computer learner corpus: A versatile new source of data for SLA research”. In S. Granger, ed., *Learner English on computer*, London: Addison Wesley Longman, pp. 3-18.

Granger, S. 1999. “Use of tenses by advanced EFL learners: evidence from an error tagged computer corpus”. *Out of corpora. Studies in Honour of Stig Johansson*. Eds. H. Hasselgard and S. Oksefjell. Amsterdam: Rodopi. 191-202.

Granger, S., Vandeventer, A. &. Hamel, M. J. 2001. “Analyse des corpus d’apprenants pour l’ELAO basé sur le TAL”. *Traitement automatique des langues* 42 (2), 609-621.

Granger, S. 2002. “A bird’s-eye view of learner corpus research”. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Eds. S. Granger, J. Hung & S. Petch-Tyson. Amsterdam: John Benjamins, 3-33.

Granger, S., Dagneaux, E. & Meunier, F. (eds.) 2002. *The International Corpus of Learner English*. Handbook and CD-ROM, Universitaires de Louvain, Louvain-la-Neuve. Available from http://www.i6doc.com

Granger, S., Hung, J. & Petch-Tyson, S. (eds) 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Granger, S. 2003a. “The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research”. To appear in *TESOL Quarterly*, special issue on corpus linguistics (Autumn 2003).

Granger, S. 2003b. “A multi-contrastive approach to the use of linkwords by advanced learners of English: evidence from the *International Corpus of Learner English.* Paper presented at the 'Pragmatic Markers in Contrast' workshop organized by the Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Brussels, 22-23 May 2003.

Granger, S. 2003c. “Error-tagged learner corpora and CALL: a promising synergy”. *CALICO Journal* 20 (3) Special issue on error analysis and error correction in computer-assisted language learning: 465-480.

Granger, S. 2004. *Centre for English Corpus Linguistics*. [Internet document available at http://cecl.fltr.ucl.ac.be/]

Hammarberg, B. 1974. “The insufficiency of error analysis”. *IRAL* 12 (3): 185-192.

Hasselgerd, H. (1999). Review of Granger (ed.), “Learner English on computer”. *ICAME Journal*, 23: 148-152.

Hutchinson, J. 1996. *UCL Error Editor*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Ide, N. & Romary, L. 2007. “Towards International Standards for Language Resources”. In L. Dybkjaer, H. Hemsen & W. Minker (eds), *Evaluation of Text and Speech Systems*. Berlin: Springer Verlag, 263-84.

Ide, N. & Sudeman, K. 2006. “Integrating Linguistic Resources: The American National Corpus Model”. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*. Genoa, Italy.

Izumi, E. et al. 2004. “SST speech corpus of Japanese learners’ English and automatic detection of learners’ errors”. *ICAME Journal* 28: 31-48. [Available online at http://nora.hd.uib.no/icame/ij28/Izumi.pdf]

Izumi, E. et al. 2005. “Error annotation for corpus of Japanese Learner English”. *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*. Jesu Island, Korea, 15 October. 71-80. [Internet document available at http://acl.ldc.upenn.edu/I/I05/I05-6009.pdf]

James, C. 1998. *Errors in Language Learning and Use. Exploring Error Analysis*. London: Longman.

Kindt, D. & Wright, M. 2001. ‘Integrating language learning and teaching with the construction of computer learner corpora”. Academia: Literature and Language. [Available from http://www.nufs.ac.jp/#kindt/media/corpora.pdf]

Källkvist, M. 1995. “Lexical errors among verbs: A pilot study of the vocabulary of advanced Swedish learners of English”. *Working papers in English and Applied Linguistics*, 2, Research Centre for English and Applied Linguistics, University of Cambridge: 103-115.

L’haire, S. & Vandeventer Faltin, A. 2003. “Error diagnosis in the FreeText project”. *CALICO Journal* 20 (3) Special issue on error analysis and error correction in computer-assisted language learning: 481-495.

Leech, G. 1993. “Corpus annotation schemes” *Literary and Linguistic Computing* 8, 4, 275–281.

Leech, G. 1997. “Introducing corpus annotation”. *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Eds. R. Garside, G. Leech and T. McEnery. London: Longman, 1-18.

Leech, G. 1998. “Learner corpora: What they are and what can be done with them”. In S. Granger, ed., *Learner English on computer*, pages xiv-xx. London: Addison Wesley Longman.

Lenko-Szymanska, A. 2003. “Lexical problems in the advanced learner corpus of written data”. Paper presented at *PALC 2003* (Practical Applications of Language Corpora), Lodz, Poland, 4-6 April 2003.

Levenston, E. A. 1971. “Over-indulgence and under-representation aspects of mother–tongue interference”. In G. Nickel (ed.), *Papers in Contrastive Linguistics*, Cambridge University Press, Cambridge.

Lüdeling, A. et al. 2005. “Multi-level error annotation in learner corpora”. *Proceedings of the Corpus Linguistics 2005 Conference.* Birmingham, United Kingdom, 14-17 July. [Internet document available at http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc]

MacWhinney, B. 2000. *The CHILDES Project, Volume 1: Tools for analysing* talk: Transcription format and programs, Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, B. 1999. “The CHILDES System”. In *Handbook of Child Language Acquisition*, Academic Press, 457-494.

Mason, O. & Uzar, R. 2000. "NLP meets TEFL: tracing the zero article". In B. Lewandowska-Tomaszczyk & P. J. Melia, eds., *PALC' 99: Practical Applications in Language Corpora*. Papers from the International Conference at the University of Lodz, pp. 105-115, Frankfurt am Main: Peter Lang.

McLaughlin, B. 1987. *Theories of second-language learning*, London: Edward Arnold.

McNeill, Arthur. 1994. "A corpus of learner errors: Making the most of a database". In L. Flowerdew & K. K. Tong (eds). *Entering Text*. Hong Kong: The Hong Kong University of Science and Technology, 114–126.

Meunier, F. 1998. "Computer tools for learner corpora". *Learner English on Computer*. Ed. S. Granger, London: Longman, 19-37.

Milton, J. & Chowdhury, N. 1994. "Tagging the interlanguage of Chinese learners of English". *Entering Text*. Eds. L. Flowerdew & K. K. Tong. Hong Kong: The Hong Kong University of Science and Technology, 127-143.

Milton, J. & Freeman, R. 1996. "Lexical variation in the writing of Chinese learners of English". In C. E. Percy, C. F. Meyer & I. Lancashire (eds). *Synchronic Corpus Linguistics. Papers from the Sixteenth International Conference on English Lanugage Research on Computerized Corpora*. Amsterdam: Rodopi, 121-131.

Milton, J. 1996. "Exploiting L1 and L2 Corpora for CALL design: The role of a hypertext grammar". In S. P. Botley, J. Glass, A. McEnery & A. Wilson (eds). *Proceeding of Teaching and Language Corpora (TALC '96)*. UCREL Technical Papers 9, Lancaster University, 233–243.

Nicholls, D. 2003. "The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT". *Proceedings of the Corpus Linguistics 2003 Conference.* Lancaster, United Kingdom, 28-31 March. 572-581.

Petch-Tyson, S. 1998. "Writer/reader visibility in EFL written discourse". In S. Granger, ed., *Learner English on computer*, pages 107-118. London: Addison Wesley Longman.

Pravec, N. A. 2002. Survey of learner corpora. *ICAME Journal*, 26: 81-114.

Scott, M. 1996. *WordSmith Tools*. Oxford University Press, Oxford.

Tan, M. 2005. "Authentic language or language errors? Lessons from a learner corpus". *ELT Journal* 59 (2): 126-134.

Tanimura, M. et al. 2004. "From learners' corpora to expert knowledge description: analyzing prepositions in the NICT JLE (Japanese Learner English) corpus". *Proceedings of IWLeL 2004: an Interactive Workshop on Language e-Learning*. Tokyo, Japan, 10 December. 139-147. [Internet document available at http://dspace.wul.waseda.ac.jp/dspace/bitstream/2065/1405/1/16.pdf]

The World Wide Web Consortium (W3C). 2006. [http://www.w3.org/XML/]

Tomaz, E. et al. 2010. “The JOS Linguistically Tagged Corpus of Slovene”. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*. Malta.

Tono, Y. 2000. “A corpus-based analysis of interlanguage development: analysing POS-tag sequences of EFL learner corpora”. In B. Lewandowska-Tomaszczyk & P. J. Melia, eds., *PALC’ 99: Practical Applications in Language Corpora*. Papers from the International Conference at the University of *Lodz*, pages 323-340, Frankfurt am Main: Peter Lang.

Tono, Y. 2003. “Learner corpora: design, development and applications”. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, United Kingdom, 28-31 March. 800-809.

Tono, Y. 1998. “Learner corpora and SLA research: Morpheme order studies revisited”. *TALC Papers*. [Available online at http://users.ox.ac.uk/~talc98/tono.htm]

Tribble, C. & Jones, G. 1997. *Concordances in the Classroom. A Resource Guide for Teachers*. Houston, Texas: Athelstan.

Tzimokas, D. 2010. “Ηλεκτρονικό σώμα κειμένων (ΗΣΚ) εκμάθησης της νέας ελληνικής ως δεύτερης γλώσσας: προς ένα ερευνητικό και διδακτικό εργαλείο”. In *Proceedings of 30th Annual Meeting of the Department of Linguistics*. Thessaloniki, 602-616.

van Rooy, B. & Schäfer, L. 2002. “The effect of learner errors on POS tag errors during automatic POS tagging”. *Southern African Linguistics and Applied Language Studies* 20: 325-335.

Van Rooy, B. & Schäfer, L. 2003. “Automatic POS tagging of a learner corpus: The influence of learner error on tagger accuracy”. In D. Archer, P. Rayson, A. Wilson & T. McEnery, eds., *Proceedings of the Corpus Linguistics 2003 Conference* (CL 2003), pp. 835-844. Lancaster University: University Centre for Computer Corpus Research on Language.