ΑΝΑΚΟΙΝΩΣΗ 34ης ΣΥΝΑΝΤΗΣΗΣ

Α. Τάντος, Κ. Αλεξάνδρη, Ι. Δοση, Κ. Πουλιου, Π. Σαββιδου & Γ. Φωτιαδου

Ανάλυση λαθών στο ελληνικό σώμα κειμένων μαθητών (ΕΣΚΕΙΜΑΘ): πρώτα ευρήματα

Abstract

This paper presents the main design principles and implementation stages of Greek Learner Corpus (GLC). GLC is a learner corpus compiled within the framework of "Education of Foreign and Repatriated Greek Students" project, funded by the European Union and the Greek Ministry of Education, Lifelong Learning and Religious Affairs. The aim of this paper is threefold: (a) to present the objectives, the current status and development of GLC; (b) to give an overview of the error annotation scheme and the complexities in the error annotation that require a stand-off annotation strategy (implemented within the annotation platform UAM Corpus Tool); and (c) to provide a first quantitative error analysis with useful hints as to the 'source' of the grammatical error and the kinds of conclusions one can derive from such an analysis in the direction of developing targeted teaching activities.

1. Εισαγωγή

Η χρησιμότητα των ηλεκτρονικών σωμάτων κειμένων στη γλωσσολογική έρευνα έγκειται στο ότι προσφέρουν στον ερευνητή τόσο τη δυνατότητα πρόσβασης σε αυθεντικό γλωσσικό υλικό, το οποίο έχει συλλεχθεί με συγκεκριμένα κριτήρια, ώστε να αποτελεί, κατά το δυνατό, αντιπροσωπευτικό δείγμα της γλώσσας (ποικιλία κειμενικών ειδών, μεγάλος όγκος δεδομένων κ.ά.), όσο και τα εργαλεία για την επεξεργασία των δεδομένων αυτών (βλ. Sinclair 1991· Leech 1992· Goutsos 2010). Μια ειδική κατηγορία σωμάτων κειμένων συνιστούν τα σώματα κειμένων μαθητικού λόγου (στο εξής ΣΚΜ). Όπως και στα γενικά σώματα κειμένων, ομοίως και στην περίπτωση των ΣΚΜ, το κύριο πλεονέκτημα της χρήσης τους έγκειται στο ότι αποτελούνται από *αυθεντικά* κείμενα παραγωγών μαθητών μιας δεύτερης/ξένης γλώσσας (Leech 1998· Hunston 2002, 15), προσφέροντας έτσι στον διδάσκοντα ή στον ερευνητή τη δυνατότητα να συναγάγει συμπεράσματα ή να αντλήσει παραδείγματα, μελετώντας την πραγματική γλώσσα των μαθητών και όχι να βασίζεται σε υποθέσεις ή σε κατασκευασμένα παραδείγματα (βλ. και Granger 1998, Leech 1998· Pravec 2002).

Στο πλαίσιο αυτό και δεδομένης της χρησιμότητας που μπορεί να έχει ένα ΣΚΜ της ελληνικής ως δεύτερης/ξένης γλώσσας, επιχειρήθηκε η δημιουργία ενός ΣΚΜ μαθητών ως μέρος του Προγράμματος "Εκπαίδευση Αλλοδαπών και Παλιννοστούντων Μαθητών" του ΑΠΘ. Στο παρόν άρθρο, παρουσιάζονται οι βασικές αρχές σχεδιασμού του "Ελληνικού Σώματος Κειμένων Μαθητών" (ΕΣΚΕΙ-

ΜΑΘ) και στη συνέχεια επιχειρείται μια πρώτη εφαρμογή του, μέσω της παρουσίασης των βασικών ευρημάτων από την ανάλυση λαθών στα δεδομένα του.

2. Ελληνικά Σώματα Κειμένων Μαθητών

Για τη νέα ελληνική έχουν γίνει ως τώρα ελάχιστες προσπάθειες στο πεδίο της δημιουργίας και της χρήσης σωμάτων κειμένων μαθητών.

Σχετική έρευνα με στόχο τη διερεύνηση των κυριότερων λαθών της διαγλώσσας παλιννοστούντων και αλλοδαπών μαθητών που φοίτησαν σε τμήματα ενισχυτικής διδασκαλίας πραγματοποιήθηκε κατά τα σχολικά έτη 2006–2007 και 2007–2008 στο πλαίσιο του προγράμματος "Ένταξη παιδιών παλιννοστούντων και αλλοδαπών στο σχολείο (ΕΠΠΑΣ) – για τη Δευτεροβάθμια Εκπαίδευση (Γυμνάσιο)" (Αναστασιάδη-Συμεωνίδη κ.ά. 2007· 2008). Η έρευνα αυτή αφορούσε γραπτές παραγωγές μαθητών όλων των επιπέδων, οι οποίες αναλύθηκαν μόνο ως προς τα *συστηματικά* λάθη.

Ωστόσο, την πρώτη συστηματική προσπάθεια κατασκευής ενός ΣΚΜ της ελληνικής αποτελεί το σώμα κειμένων γλωσσικής εκμάθησης που σχεδιάστηκε αρχικά στο πλαίσιο του Ερευνητικού Προγράμματος "Πυθαγόρας Ι" του Πανεπιστημίου Αθηνών (2004) και η ανάπτυξή του συνεχίζεται ως σήμερα (βλ. Τζιμώκας 2010). Στο πλαίσιο της δημιουργίας του συγκεκριμένου ΣΚΜ γίνεται μια ενδελεχής προσπάθεια κωδικοποίησης μιας αντιπροσωπευτικής ποικιλίας λαθών από ενήλικες μαθητές της ελληνικής ως δεύτερης/ξένης γλώσσας. Συγκεκριμένα, πρόκειται για ένα σώμα κειμένων που αποτελείται από περίπου 65.000 λέξεις, οι οποίες αντιστοιχούν σε 191 παραγωγές γραπτού λόγου ενήλικων μαθητών διαφόρων επιπέδων γλωσσομάθειας. Ένα από τα πλεονεκτήματά του αποτελεί και η ποικιλία των κειμενικών ειδών που περιλαμβάνει. Επιπλέον, το συγκεκριμένο ΣΚΜ βασίζεται σε ένα πολύ λεπτομερές πλαίσιο επισημείωσης λαθών, το οποίο βοηθά στην ολοκληρωμένη κατηγοριοποίηση των λαθών. Εντούτοις, ένα μειονέκτημα είναι ότι ενίοτε το πλαίσιο επισημείωσης κατευθύνει τον εκάστοτε επισημειωτή στην υποκειμενική ερμηνεία της φύσης του λάθους, περιορίζοντας με αυτό τον τρόπο την ελευθερία των μελλοντικών χρηστών του (ερευνητών, διδασκόντων) ως προς την ερμηνεία των λαθών. Έτσι, παρά το εντυπωσιακό εύρος δεδομένων και τη συστηματικότητα του συγκεκριμένου ΣΚΜ, το πλαίσιο επισημείωσής του διακρίνεται από έλλειψη ευελιξίας, περιορίζοντας την εμβέλεια χρήσης και ανάλυσης των δεδομένων του και καθιστά δύσκολο τον παραλληλισμό του με ΣΚΜ άλλων γλωσσών.

3. Το Ελληνικό Σώμα Κειμένων Μαθητών (ΕΣΚΕΙΜΑΘ)

Η κατασκευή του ΕΣΚΕΙΜΑΘ βασίστηκε στην αρχή της Αντιπαραβολικής Ανάλυσης της Διαγλώσσας (Contrastive Interlanguage Analysis, CIA· βλ. Granger 1998), η οποία στηρίζεται στην αντιπαραβολή της διαγλώσσας σε δύο επίπεδα: (α) τη

σύγκριση της διαγλώσσας με τη γλώσσα-στόχο και (β) τη σύγκριση ανάμεσα σε διαγλώσσες διαφορετικών ατόμων (με κοινή γλώσσα-στόχο). Η βάση υλοποίησης αυτής της αρχής υπήρξε η ταξινόμηση, επισημείωση και ανάλυση των λαθών των μαθητών με διαφορετική πρώτη γλώσσα. Αντίστοιχα, το ΕΣΚΕΙΜΑΘ ακολούθησε τα εξής στάδια:

(1) συλλογή δεδομένων·

(2) καταγραφή των βασικών κατηγοριών των γραμματικών λαθών των μαθητών πάνω σε δείγμα 50 κειμένων·

(3) σχεδιασμός του πλαισίου επισημείωσης των λαθών·

(4) επισημείωση κειμένων με το εργαλείο επισημείωσης UAM Corpus Tool (βλ. 3.2.1) με ταυτόχρονες συχνές αναθεωρήσεις του πλαισίου επισημείωσης·

(5) διασταύρωση και τυποποίηση της στρατηγικής επισημείωσης.

Στις επόμενες ενότητες ακολουθεί η περιγραφή των παραπάνω σταδίων υλοποίησης του ΕΣΚΕΙΜΑΘ καθώς και η ποσοτική ανάλυση λαθών και εξαγωγή των πρώτων συμπερασμάτων για την ανίχνευση τάσεων στις παραγωγές των μαθητών.

3.1 Δεδομένα ΕΣΚΕΙΜΑΘ

Στην πρώτη του έκδοση, το ΕΣΚΕΙΜΑΘ αποτελείται από περίπου 500(+) κείμενα και ο αριθμός λέξεών του ανέρχεται περίπου στις 33.500. Τα δεδομένα του έγκεινται σε γραπτές παραγωγές στη βάση σταθμισμένων τεστ γλωσσομάθειας και προέρχονται από μαθητές που φοιτούν στις τάξεις υποδοχής της πρωτοβάθμιας και δευτεροβάθμιας εκπαίδευσης. Τα δεδομένα του ΕΣΚΕΙΜΑΘ προέρχονται από γραπτές παραγωγές 329 μαθητών (αγόρια, N=211· κορίτσια, N=118) που φοιτούσαν σε 25 σχολεία της πρωτοβάθμιας και της δευτεροβάθμιας εκπαίδευσης από διάφορες περιφέρειες της Ελλάδας (Αττική, Θεσσαλία, Κρήτη κ.α.) κατά το σχολικό έτος 2011–2012. Όσον αφορά τη μητρική τους γλώσσα οι περισσότεροι (σε ποσοστό 46,8%) είναι ομιλητές της αλβανικής ενώ το 73,3% δεν έχουν καμία δεύτερη μητρική γλώσσα. Χώρες γέννησης των συμμετεχόντων είναι κατά κύριο λόγο η Ελλάδα (σε ποσοστό 50%) και η Αλβανία (σε ποσοστό 32,2%) ενώ οι περισσότεροι από αυτούς έχουν χώρα καταγωγής την Αλβανία (48,3%), τη Ρωσία (12,8) και τη Γεωργία (12,1%).

Στους συγκεκριμένους μαθητές χορηγήθηκαν τα επικαιροποιημένα τεστ ελληνομάθειας "Ας μιλήσουμε Ελληνικά" I, II και III, στα οποία περιλαμβάνονται τρεις προσχεδιασμένες δοκιμασίες παραγωγής γραπτού λόγου. Οι δοκιμασίες εξετάζουν το κειμενικό είδος της αφήγησης και χορηγήθηκαν σε παιδιά διαφορετικών ηλικιών. Συγκεκριμένα, το τεστ ελληνομάθειας I (Τεστ 1 στο εξής) έχει χορηγηθεί σε παιδιά Α΄ και Β΄ Δημοτικού, το τεστ ελληνομάθειας II (Τεστ 2 στο εξής) έχει χορηγηθεί σε παιδιά Γ΄ και Δ΄ Δημοτικού ενώ το τεστ ελληνομάθειας III χορηγήθηκε σε παιδιά Ε΄ και Στ΄ Δημοτικού (Τεστ 3 Α΄ βάθμιας στο εξής) αλλά και στη δευτεροβάθμια εκπαίδευση σε μαθητές Γυμνασίου (Τεστ 3 Β΄ βάθμιας στο εξής).

3.2 Τεχνικές και πλαίσιο επισημείωσης

Ένα από τα πιο σημαντικά στάδια στον σχεδιασμό ενός ΣΚΜ είναι ο καθορισμός του πλαισίου επισημείωσης των λαθών.[1] Στην κατασκευή του ΕΣΚΕΙΜΑΘ, μια βασική αρχή που καθοδήγησε τις επιλογές όσον αφορά τόσο τον τεχνικό σχεδιασμό του (λογισμικό επισημείωσης και πρακτικές) όσο και την κατηγοριοποίηση των λαθών (καθορισμός κατηγοριών και υποκατηγοριών λαθών, π.χ. ορθογραφικά κτλ.) ήταν η επιδιωκόμενη *ευελιξία* του τελικού προϊόντος, του ΕΣΚΕΙΜΑΘ, ως εργαλείου που μπορεί να χρησιμοποιηθεί στο μέλλον τόσο από ερευνητές όσο και από διδάσκοντες/διδασκομένους.

3.2.1 Τεχνικές επισημείωσης

Μεθοδολογικά υιοθετήθηκαν τα πρότυπα του προγράμματος Free Text για το γαλλικό ΣΚΜ FRIDA του 1998, το οποίο ακόμη αποτελεί διεθνώς σημείο αναφοράς για τα ΣΚΜ. Πιο συγκεκριμένα, ακολουθώντας τις κύριες μεθοδολογικές αρχές του Free Text, η δημιουργία και ανάλυση του ΕΣΚΕΙΜΑΘ βασίστηκε στα ακόλουθα βήματα: (α) χειρωνακτική εύρεση των λαθών· (β) επεξεργασία και συμφωνία για το σύνολο των ετικετών που θα απαρτίζουν το πλαίσιο επισημείωσης λαθών· (γ) εισαγωγή ετικετών λαθών και διορθώσεων στα αρχεία κειμένων· (δ) ανάκτηση/εξαγωγή λιστών συγκεκριμένων τύπων λαθών και στατιστική ανάλυσή τους· (ε) γλωσσική ανάλυση των σημαντικότερων τύπων λαθών, βάσει κυρίως συμφραστικών πινάκων (concordance-based).

Όσον αφορά το τεχνικό μέρος, ακολουθώντας ως πρότυπο την Granger (2003), το πλαίσιο επισημείωσης των λαθών του ΕΣΚΕΙΜΑΘ σχεδιάστηκε με στόχο να είναι: (α) *διαφωτιστικό* και ταυτόχρονα *διαχειρίσιμο*: να είναι αναλυτικό σε βαθμό που να παρέχει χρήσιμες πληροφορίες για τα λάθη των μαθητών και ταυτόχρονα να μπορεί να το διαχειριστεί ο επισημειωτής αλλά και ο τελικός χρήστης· (β) *επαναχρησιμοποιήσιμο*: οι κατηγορίες λαθών να είναι αρκετά γενικές, ώστε να μπορούν να χρησιμοποιηθούν για διαφορετικές γλώσσες· (γ) *ευέλικτο*: να επιτρέπει άμεση πρόσβαση για αλλαγές (προσθήκη/αφαίρεση ετικετών) πάνω στα επισημειωμένα κείμενα· (δ) *συνεπές*: να μην υπάρχουν αντιφάσεις στις επισημειώσεις των κειμένων, όταν αναμειγνύονται περισσότεροι του ενός επισημειωτές.

Ένα από τα πιο σημαντικά πλεονεκτήματα του ΕΣΚΕΙΜΑΘ σε σχέση με προηγούμενα ΣΚΜ για την ελληνική είναι ότι εφαρμόζεται η λεγόμενη *εξ αποστάσεως* (stand-off) επισημείωση, στην οποία οι επισημειώσεις διαχωρίζονται από τα αυθεντικά ψηφιοποιημένα κείμενα. Η εξ αποστάσεως επισημείωση προσφέρει τη δυνατότητα πολλαπλών επισημειώσεων στα ίδια δεδομένα, είναι ευέλικτη και επεκτάσιμη, ενώ το ότι επιτρέπει την αποθήκευση των επισημειώσε-

[1] Για την πρακτική αξιοποίηση των επισημειωμένων κειμένων, βλ. ενδεικτικά Leech (1998) και Granger (1998· 2002), όπου τίθενται ποικίλα μεθοδολογικά ζητήματα. Για ειδικότερες αναφορές σε εργαλεία και λογισμικά επεξεργασίας ΣΚΜ, βλ. Meunier (1998).

ων και των αυθεντικών κειμένων σε ξεχωριστά αρχεία, καθιστά δυνατή τη χρήση του ΣΚΜ με τυχόν διαφορετικό πλαίσιο επισημείωσης λαθών από μελλοντικούς ερευνητές. Έτσι επιτυγχάνεται η ευελιξία στη χρήση του ΣΚΜ και η δυνατότητα προσαρμογής του στους στόχους της εκάστοτε έρευνας. Αξίζει να επισημανθεί πως το ΕΣΚΕΙΜΑΘ είναι ένα από τα πρώτα ΣΚΜ διεθνώς που ακολουθούν τη εξ αποστάσεως επισημείωση.

Η επισημείωση των κειμένων πραγματοποιείται με το λογισμικό UAM Corpus Tool (O' Donnell 2008), το οποίο υποστηρίζει την εξ αποστάσεως επισημείωση, καθώς επιτρέπει τη χρήση πολλαπλών πλαισίων επισημείωσης, αποθηκεύοντας τα επισημειωμένα κείμενα σε ξεχωριστά αρχεία για το κάθε πλαίσιο επισημείωσης.

Επιπλέον, το ΕΣΚΕΙΜΑΘ έχει επισημειωθεί ως προς τα μέρη του λόγου[2] και η επιλογή της εξ αποστάσεως επισημείωσης επιτρέπει τη χρήση των δεδομένων του σε τέσσερις διαφορετικές μορφές:

- ως μη επισημειωμένο σώμα κειμένων (χρήση των αυθεντικών ψηφιοποιημένων κειμένων, χωρίς τις επισημειώσεις)·
- ως επισημειωμένο ως προς τα λάθη·
- ως επισημειωμενο ως προς τα μέρη του λόγου·
- ως επισημειωμένο τόσο ως προς τα λάθη όσο και ως προς τα μέρη του λόγου.

3.2.2 Πλαίσιο επισημείωσης λαθών

Σχετικά με την κατηγοριοποίηση των λαθών, υιοθετήθηκε αρχικά ο πρακτικός διαχωρισμός των Dulay, Burt & Krashen (1982) σε λάθη ορισμένα με βάση: (α) γλωσσολογικές κατηγορίες (όπως χρόνος, όψη, γένος) και (β) τον τρόπο πραγμάτωσής τους (παράλειψη, προσθήκη, αντικατάσταση κτλ.). Η κατηγοριοποίηση γίνεται ιεραρχικά και περιλαμβάνει τρία στρώματα ιεράρχησης: τομέας λαθών (error domain)· κατηγορία λαθών (error category)· γραμματική κατηγορία της λέξης (word category).

Η κατηγοριοποίηση των λαθών του ΕΣΚΕΙΜΑΘ περιγράφεται αναλυτικά στον Πίνακα 1. Ο καθορισμός των κατηγοριών διαφοροποιείται από τον αντίστοιχο των ΣΚΜ άλλων γλωσσών, καθώς για τον σχεδιασμό του λήφθηκε υπόψη η πλούσια μορφολογία της ελληνικής. Η τελική μορφή του πλαισίου επισημείωσης (αριθμός και είδος κατηγοριών) προέκυψε μετά από ενδελεχή μελέτη και συνεχείς αναθεωρήσεις από τα μέλη της ερευνητικής ομάδας του ΕΣΚΕΙΜΑΘ. Οι αναθεωρήσεις καθοδηγούνταν από τα ίδια τα δεδομένα, καθώς ένα από τα στάδια της κατασκευής του ΕΣΚΕΙΜΑΘ ήταν η δοκιμαστική επισημείωση αντιπροσωπευτικού μέρους των δεδομένων του με στόχο να ελεγχθεί και

[2] Το ΕΣΚΕΙΜΑΘ, στην παρούσα μορφή του, έχει επισημειωθεί μόνο τμηματικά, αλλά στην τελική του μορφή θα είναι πλήρως επισημειωμένο ως προς τα μέρη του λόγου. Λόγω περιορισμού έκτασης, στην παρούσα εργασία δεν γίνεται εκτενέστερη αναφορά στην επισημείωση των μερών του λόγου. Για έναν προβληματισμό σχετικά με την αξιοπιστία της επισημείωσης των μερών του λόγου σε ΣΚΜ, βλ. Van Rooy & Schöfer (2003). Για μια εξέταση της σκοπιμότητας της συντακτικής ανάλυσης (parsing) των ΣΚΜ, βλ. de Mönnink (2000).

να αναθεωρηθεί η αρχική κατηγοριοποίηση των λαθών. Εκτός από τα είδη των λαθών που εντοπίστηκαν στα δεδομένα (τα οποία έχουν να κάνουν τόσο με τον ιδιαίτερο χαρακτήρα της ελληνικής όσο και με τα χαρακτηριστικά των συγκεκριμένων κειμένων), ο καθορισμός του πλαισίου επισημείωσης του ΕΣΚΕΙΜΑΘ υπαγορεύτηκε από τον βασικό στόχο του, ο οποίος έγκειται στο να αποτυπώσει την εικόνα των εξωτερικών χαρακτηριστικών της διαγλώσσας των μαθητών, χωρίς να μπει σε διαδικασία ερμηνείας. Επιλέχθηκε, επομένως, μια περιγραφική κατηγοριοποίηση, έναντι μιας ταξινόμησης που θα υπεισερχόταν σε διαδικασία ερμηνείας των λαθών.

Οι γραπτές παραγωγές των μαθητών έχουν επισημειωθεί με βάση το ακόλουθο πλαίσιο επισημείωσης στο οποίο διακρίνονται οι κατηγορίες και οι υποκατηγορίες των λαθών:

Τομέας Λαθών		**Κατηγορία Λαθών**			
Είδος	**Περιγραφή**	**Είδος**	**Περιγραφή ΕΣΚΕΙΜΑΘ**		
ΟΡΘ	Ορθογραφία	-ΛΕΞ	Λεξικό μόρφημα		
		-ΓΡΑΜ	Κλιτικό–γραμματικό μόρφημα		
		-ΠΑΡ	Παράλειψη		
		-ΚΕΦ	Κεφαλαία		
		-ΠΕΖ	Πεζά		
		-ΚΩΔ	Κώδικας γραφής		
ΓΡΑΦ	Γραφηματικά	-ΛΕΞ	Λεξικό μόρφημα		
		-ΓΡΑΜ	Γραμματικό μόρφημα		
ΤΟΝ	Τονισμός	-ΠΡΟΣ	Προσθήκη		
		-ΠΑΡ	Παράλειψη		
		-ΑΝΤΙ	Αντικατάσταση		
ΣΥΜΦ	Συμφωνία	-ΑΡΙ	Αριθμός		
		-ΓΕΝ	Γένος		
		-ΠΡΟΣ	Πρόσωπο		
		-ΠΤΩ	Πτώση		
ΑΡΘ	Άρθρα	ΟΡΙ	Οριστικό	-ΠΑΡ	Παράλειψη
				-ΠΡΟΣ	Προσθήκη
				-ΑΝΤΙ	Αντι/ση
		ΑΟΡ	Αόριστο	-ΠΑΡ	Παράλειψη
				-ΠΡΟΣ	Προσθήκη
				-ΑΝΤΙ	Αντι/ση
ΚΛΙΤ	Κλιτικά	-ΠΑΡ	Παράλειψη		
		-ΠΡΟΣ	Προσθήκη		
		-ΑΝΤΙ	Αντικατάσταση		

Τομέας Λαθών		**Κατηγορία Λαθών**			
Είδος	**Περιγραφή**	**Είδος**	**Περιγραφή ΕΣΚΕΙΜΑΘ**		
ΣΔ	Συμπληρωματικοί δείκτες	-ΠΑΡ	Παράλειψη		
		-ΠΡΟΣ	Προσθήκη		
		-ΑΝΤΙ	Αντικατάσταση		
ΓΕΝ	Γένος		Απόδοση γένους		
ΧΡ	Χρόνος	-ΠΡΛΘ	Παρελθοντικός τύπος		
		-ΜΠΡΛΘ	Μη παρελθοντικός τύπος		
		-ΠΡΟΣ	Προσθήκη		
		-ΑΝΤΙ	Αντικατάσταση		
ΟΨΗ	Γραμματική όψη	-ΣΥΝ	Συνοπτικός τύπος		
		-ΜΣΥΝ	Μη συνοπτικός τύπος		
ΦΩΝΗ	Φωνή	-ΕΝΕΡΓ	Ενεργητικός τύπος		
		-ΠΑΘ	Παθητικός τύπος		
ΣΟΡ	Σειρά όρων				
ΛΕΞΗ	Λεξικά	-ΠΑΡ	Παράλειψη		
		-ΠΡΟΣ	Προσθήκη		
ΚΕΙΜ	Κειμενικά (π.χ. συνοχή)	-ΣΥΝΔ	Σύνδεσμοι	-ΠΑΡ	Παράλειψη
				-ΠΡΟΣ	Προσθήκη
				-ΑΝΤΙ	Αντι/ση
ΣΤΙΞ	Στίξη	-ΠΑΡ	Παράλειψη		
Ν	Τελικό *-ν*				
ΠΡΟΒΛ	Προβληματικά				

Πίνακας 1: Πλαίσιο επισημείωσης – επεξήγηση ονομασίας επισημειώσεων

Ενδεικτικά παρατίθενται τα παρακάτω παραδείγματα επισημείωσης:

[1] αρχίσαμε να *παίξουμε
Ετικέτα λάθους: _ΟΨΗ_ΣΥΝ

[2] *ένας περιστέρι
Ετικέτα λάθους: (α) _ΓΕΝ (β) _ΣΥΜΦ_ΓΕΝ

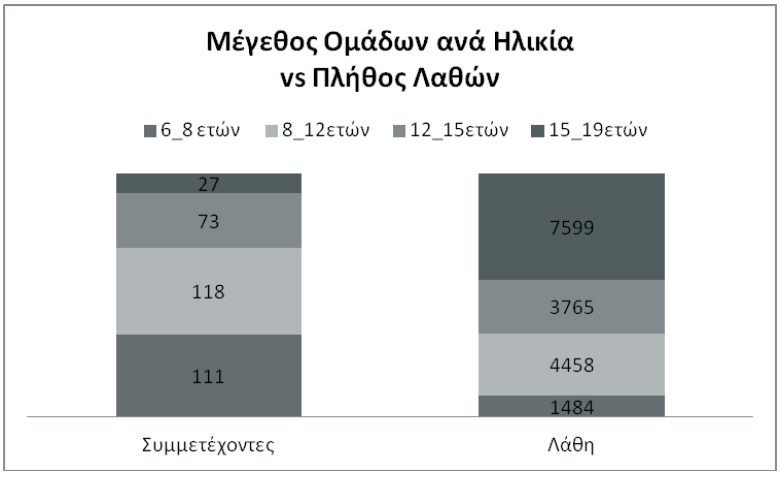
[3] *μία δέντρα
Ετικέτα λάθους: (α) _ΓΕΝ (β)_ΣΥΜΦ_ΓΕΝ/ΑΡΙ

Στο παράδειγμα [1] χρησιμοποιείται ο συνοπτικός τύπος του ρήματος *παίζω* αντί του μη συνοπτικού. Στην περίπτωση αυτή η ετικέτα λάθους είναι 'όψη' ως γενική κατηγορία λάθους και 'συνοπτικό' ως υποκατηγορία. Σε περιπτώσεις όπως τα παραδείγματα [2] και [3] υπάρχει η δυνατότητα πολλαπλής επισημείωσης. Συγκεκριμένα, στο [2] το λάθος αποδίδεται είτε ως προς την συμφωνία του γένους (ΣΥΜΦ_ΓΕΝ) ανάμεσα στον προσδιοριστή και το ουσιαστικό είτε

ως προς την απόδοση του γένους του ουσιαστικού (ΓΕΝ). Αντίστοιχη είναι και η τελευταία περίπτωση (παράδειγμα [3]) στην οποία όμως προστίθεται και η δυνατότητα επισημείωσης για την πιθανότητα λάθους συμφωνίας ως προς τον αριθμό (ΣΥΜΦ_ΑΡΙ) ανάμεσα στον προσδιοριστή και στο ουσιαστικό. Οι επισημειώσεις των γραπτών παραγωγών γίνονται στο περιβάλλον UAM Corpus Tool (O' Donnell 2008). Ταυτόχρονα το ίδιο κείμενο υπάρχει σε τρεις μορφές: αυθεντικό, επισημειωμένο ως προς τα λάθη και επισημειωμένο ως προς τα μέρη του λόγου.

3.2.3 Ανάλυση των λαθών

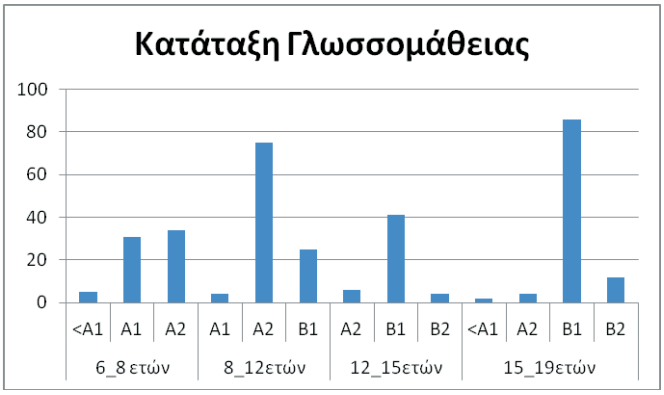
Έχοντας συγκεντρώσει και επισημειώσει τις γραπτές παραγωγές των μαθητών με βάση το πλαίσιο επισημείωσης λαθών που προαναφέρθηκε και προχωρώντας σε μια πρώτη ανάλυση των λαθών, εξετάζουμε την αντιστοιχία μεταξύ του πλήθους των παραγωγών ανά ηλικιακή ομάδα σε σχέση με το πλήθος των λαθών που ανιχνεύθηκαν (Γράφημα 1).



Γράφημα 1: Μέγεθος ομάδων ανά ηλικία vs πλήθος λαθών

Όπως φαίνεται στο παραπάνω γράφημα, δεν υπάρχει αναλογία ένα προς ένα ανάμεσα στον αριθμό των συμμετεχόντων στις διάφορες ηλικιακές ομάδες και στο πλήθος λαθών που παρήγαγαν. Ενώ η πλειοψηφία των παραγωγών που αναλύθηκαν προέρχεται από παιδιά 6–8 ετών και 8–12 ετών (Τεστ 1 και 2 αντίστοιχα), η πλειοψηφία των λαθών συγκεντρώνεται στα παιδιά μεγαλύτερης ηλικίας (Τεστ 3 Β΄ βάθμιας).

Στη συνέχεια εξετάζουμε τη συσχέτιση μεταξύ του πλήθους των παραγωγών ανά ηλικιακή ομάδα με την κατάταξη σε συγκεκριμένα επίπεδα γλωσσομάθειας (Γράφημα 2).



Γράφημα 2: Κατάταξη επιπέδων γλωσσομάθειας ανά ηλικία

Όπως φαίνεται τα παιδιά μικρότερων ηλικιών είναι μοιρασμένα ανάμεσα στο Α1 και το Α2 επίπεδο, αν και έχουν παρακολουθήσει όλα τα χρόνια εκπαίδευσης το ελληνικό σύστημα. Η ομάδα των 9–12 ετών είναι κυρίως επιπέδου Α2, ενώ η ομάδα των 12–15 ετών είναι Β1. Τα παιδιά μεγαλύτερης ηλικίας έχουν μεγαλύτερες παραγωγές και η πλειονότητά τους κατατάσσεται στο Β1 επίπεδο γλωσσομάθειας.

Προχωρώντας σε μια αναλυτική περιγραφή της κατανομής των λαθών στην κάθε ομάδα παρατηρούμε πως αυτή δεν διαφοροποιείται σε μεγάλο βαθμό. Ο Πίνακας 2 παρουσιάζει τις συχνότητες από όλα τα λάθη που βρέθηκαν στις επισημειωμένες παραγωγές.

			Τεστ 1	Τεστ 2	Τεστ 3 Α΄βάθμιας	Τεστ 3 Β΄βάθμιας
άρθρο	αόριστο	παράλειψη	7	1	1	6
		αντικατάσταση				1
	οριστικό	παράλειψη	2		2	8
		προσθήκη	1		5	10
		αντικατάσταση		4	4	25
γένος			11	18	15	97
συμφωνία	αριθμός		14	25	15	53
	γένος		9	17	15	109
	πρόσωπο		3	14	9	22
	πτώση		9	21	27	56
λέξη		αντικατάσταση	16	29	32	71
		παράλειψη	14	35	31	57
		προσθήκη	9	38	23	51
χρόνος	μη-παρελθόν		12	15	21	107
	παρελθόν			4	3	5

			Τεστ 1	Τεστ 2	Τεστ 3 Α΄βάθμιας	Τεστ 3 Β΄βάθμιας
όψη	μη-συνοπτικό		21		31	65
	συνοπτικό		1	4	2	16
ορθογραφία	γραμματικό		227	422	287	566
	τόνος		1	6	7	16
	κεφαλαίο		19	48	32	62
	μικρό		12	46	51	73
	λεξικό		298	1005	732	1263
	λατινικό					47
	σύμπτυξη		51	141	74	155
	τελικό *-ν*		2		20	93
	τόνος	αντικατάσταση	29	147	106	174
		παράλειψη	389	1553	1561	3105
		προσθήκη	6	8	30	66
γραφηματικό	γραμματικό		52	53	57	67
	λεξικό		90	274	148	314
στίξη		αντικατάσταση	4	22	17	75
		παράλειψη	63	307	270	540
		προσθήκη	10	14	11	48
συλλαβή						3
κειμενικό			77	119	74	108
ΣΟΡ			2	6	9	18
προβληματικό			23	62	43	47

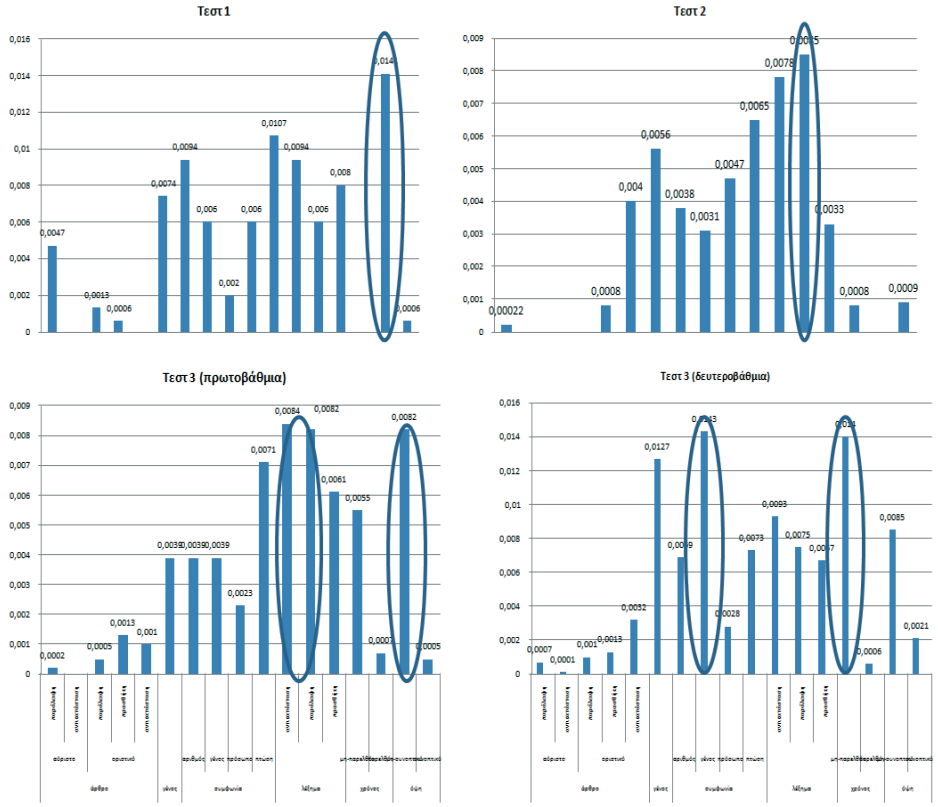
Πίνακας 2: Κατανομή λαθών στα Τεστ 1, 2, 3 (Α΄ βάθμιας, Β΄ βάθμιας)

Όπως φαίνεται στον Πίνακα 2, το πιο συχνό λάθος αποτελεί η *παράλειψη τόνου*. Οι διαφορές ανάμεσα στα Τεστ 1 και 2, όπως και στα Τεστ 2 και 3 είναι στατιστικά σημαντικές (χ^2=152.179, p<.001, η=.175, φ=−.175 και χ^2=8.745, p=.003, η=.035, φ=−.035), ενώ δεν σημειώνονται σημαντικές διαφορές στις παραγωγές στο Τεστ 3 (Α΄ βάθμιας και Β΄ βάθμιας). Η έλλειψη λαθών στο Τεστ 1 μπορεί να οφείλεται στη μικρή έκταση των παραγωγών και στη συχνή χρήση άτονων λέξεων.

Πολύ συχνά είναι επίσης τα *ορθογραφικά λάθη*, τόσο στο *λεξικό* όσο και στο *γραμματικό* μόρφημα (βλ. Πίνακα 2). Τα λάθη στο *λεξικό* μόρφημα διαφοροποιούνται κυρίως ανάλογα με το Τεστ και λιγότερο με βάση την ηλικία: Τεστ 1 vs 2: χ^2=38.900, p<.001, η=.081, φ=−.081· Τεστ 2 vs 3: χ^2=8.802, p=.003, η=.033, φ=−.033· ενώ Τεστ 3 (Α΄ βάθμιας) vs Τεστ 3 (Β΄ βάθμιας): μη σημαντικές διαφορές. Από την άλλη, τα ορθογραφικά λάθη στο *γραμματικό* μόρφημα μειώνονται πιο

έντονα μετά την ηλικία των 8 ετών (> Γ΄ Δημοτικού, Τεστ 2 και 3): Τεστ 1 vs 2: χ^2=3.945, p=.047, η=.026, φ=−.026· Τεστ 2 vs 3: χ^2=11.784, p=.001, η=.038, φ=−.038· και Τεστ 3 (Α΄ βάθμιας) vs (Β΄ βάθμιας): χ^2=13.849, p<.001, η=.035, φ=−.035.

Περνώντας στα *γραφηματικά λάθη* παρατηρούμε ότι τα λάθη στο *γραμματικό* μόρφημα μειώνονται αναπτυξιακά και τείνουν να εξαφανιστούν, ενώ στο λεξικό μόρφημα παραμένουν, αν και μειώνονται και αυτά σημαντικά (Τεστ 1 vs Τεστ 3 Β΄ βάθμιας: χ^2=6326.784, p<.001, η=.085, φ=.085· και Τεστ 1 vs Τεστ 3 Β΄ βάθμιας: χ^2=10.910, p=.001, η=.035, φ=.035, αντίστοιχα).



Γράφημα 3: Κατανομή (%) μορφο-συντακτικών λαθών ανά τεστ/ηλικία

Τα *γραμματικά (μορφο-συντακτικά) λάθη* δεν έχουν μεγάλη συχνότητα, αλλά αξίζει μια περαιτέρω ανάλυσή τους. Η κατανομή της συχνότητας εμφάνισής τους ποικίλλει ανάλογα με το τεστ/ηλικία. Το Γράφημα 3 δείχνει ότι το πιο συχνό λάθος στο Τεστ 1 είναι η χρήση μη-συνοπτικής όψης (βλ. Πίνακα 2). Στα Τεστ 2 και 3 (Α΄ βάθμιας), είναι πολύ συχνά τα λάθη *παράλειψη*, *προσθήκη* και *αντικατάσταση λέξης*, κατηγορία που περιέχει τη χρήση των κλιτικών, της μορφολογίας της φωνής και άλλων, γι' αυτό δεν κατατίθενται σχετικά συμπερά-

σματα. Στα δύο αυτά τεστ, ακολουθούν σε συχνότητα τα λάθη *συμφωνίας* στον *αριθμό*, το *γένος* και την *πτώση*, ενώ στο Τεστ 3 (Β΄ βάθμιας) μεγάλη συχνότητα έχουν τα λάθη *συμφωνίας* ως προς το *γένος* καθώς και η λανθασμένη χρήση *μη-παρελθοντικών χρόνων*.

Τα λάθη συμφωνίας *γένους* παρουσιάζουν σταθερή εμφάνιση στις παραγωγές, εκτός του Τεστ 3, όπου είναι σημαντικά περισσότερες (Τεστ 3 Α΄ βάθμιας vs Β΄ βάθμιας: χ^2=25.037, p<.001, η=.047, φ=−.047). Τα λάθη συμφωνίας *αριθμού*, *προσώπου* και *πτώσης* παραμένουν σταθερά σε χαμηλά επίπεδα, χωρίς σημαντικές διαφορές. Τα πορίσματα της παρούσας ανάλυσης δείχνουν ότι τα προβλήματα στη χρήση του γένους είναι εντονότερα σε περιβάλλοντα φυσικής κατάκτησης (Dimitrakopoulou κ.ά. 2004· 2006· 2007).

Περνώντας στη ρηματική μορφολογία, τα λάθη χρήσης *όψης* μειώνονται, αν και όχι σημαντικά (Τεστ 1 vs Τεστ 3 Α΄ βάθμιας: χ^2=3.800, p=.051, η=.027, φ=−.027). Ανάλογα προβλήματα εντοπίζονται σε προχωρημένα επίπεδα γλωσσομάθειας (Παπαδοπούλου 2005· Tsimpli & Papadopoulou 2006). Από την άλλη, τα λάθη στη χρήση *χρόνου* εμφανίζονται σημαντικά συχνότερα στα παιδιά μεγαλύτερης ηλικίας (Τεστ 3 Α΄ βάθμιας vs Τεστ 3 Β΄ βάθμιας: χ^2=16.345, p<.001, η=.038, φ=−.038). Ωστόσο, δεν μπορούμε να εκτιμήσουμε την εμφάνιση λαθών στη χρήση *μη-παρελθοντικών* χρόνων σε αυτό τον πληθυσμό, καθώς χρειάζεται έλεγχος πιθανότητας συσχέτισης του πλήθους των λαθών με την ηλικία πρώτης επαφής με την ελληνική και τα συνολικά έτη έκθεσης στη γλώσσα και στο ελληνικό εκπαιδευτικό σύστημα.

Τέλος, τα λάθη σε *κειμενικούς δείκτες* και *συνοχή*, όπως και στη *σειρά των όρων της πρότασης* (Πίνακας 2) έχουν χαμηλή συχνότητα εμφάνισης. Τα παραπάνω είναι πιθανόν να οφείλονται στη μεθοδολογία επισημείωσης: η έλλειψη *συνοχής* σε ολόκληρο κείμενο εμφανίζεται ως ένα λάθος, χωρίς αυτό να συνεπάγεται ομοιομορφία ανάμεσα στα διαφορετικά λάθη που εντοπίζονται στα κείμενα που αναλύθηκαν.

Συνολικά, τα ευρήματα συμφωνούν με προηγούμενες έρευνες (Αναστασιάδη-Συμεωνίδη κ.ά. 2007· 2008), καθώς η κατανομή της συχνότητας των λαθών εμφανίζει ομοιογενή εικόνα στους πληθυσμούς δίγλωσσων παιδιών που έχουν εξεταστεί.

4. Συμπεράσματα και μελλοντική ανάπτυξη του ΕΣΚΕΙΜΑΘ

Το ΕΣΚΕΙΜΑΘ είναι ένα εύχρηστο εργαλείο τόσο για εξειδικευμένο προσωπικό (ερευνητές, εκπαιδευτικούς) όσο και για τους μαθητές. Τα δεδομένα που συλλέγονται με βάση τις αυθεντικές παραγωγές των συμμετεχόντων παρέχουν χρήσιμες πληροφορίες για την εξαγωγή συμπερασμάτων σε σχέση με τη διαγλώσσα των συγκεκριμένων μαθητών. Μάλιστα, η ποσοτική ανάλυση της εμφάνισης λαθών, μπορεί να αποδειχθεί ιδιαίτερα ευεργετική για τη βελτίωση του πλαισίου κατάταξης σε επίπεδα γλωσσομάθειας. Έτσι, για παράδειγμα, συμπεράσματα για λάθη που εμμένουν και σε προχωρημένα επίπεδα γλωσσομάθειας (π.χ. Β1),

όπως η χρήση του γένους, της όψης και του χρόνου, μπορούν να λειτουργήσουν καθοδηγητικά για τη διαμόρφωση αναλυτικών προγραμμάτων σπουδών και γλωσσοδιδακτικών μεθόδων με σκοπό τη βελτίωση της επίδοσης των μαθητών.

Βιβλιογραφία

Αναστασιάδη-Συμεωνίδη, Ά., Ε. Βλέτση, Μ. Μητσιάκη, Β. Μποζονέλος & Β. Χούμα. 2008. "Τα γλωσσικά λάθη των μαθητών της ελληνικής ως δεύτερης γλώσσας και ο ρόλος της Γι στις πολυπολιτισμικές τάξεις του Γυμνασίου", στο *International Conference "European Year of Intercultural Dialogue: Discussing with Languages-Cultures"*. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 597–612. [http://www.frl.auth.gr/sites/congres/Interventions/GR/anastassiadis-symeonidis.pdf]

Αναστασιάδη-Συμεωνίδη Ά., Α. Ζουραβλιόβα, Μ. Μητσιάκη & Γ. Φωτιάδου. 2007. "Τα φωνολογικά λάθη στη διαγλώσσα των σλαβόφωνων μαθητών της ελληνικής ως δεύτερης/ξένης γλώσσας", στο *Διεθνές συνέδριο της ΕΕΕΓ*. Θεσσαλονίκη. [http://www.enl.auth.gr/gala/14th/Papers/Greek%20papers/Anastasiadi-Symeonidi&Zouravliova&Mitsiaki&Fotiadou.pdf]

Corder, S. P. 1981. *Error Analysis and Interlanguage*. Οξφόρδη: Oxford University Press.

Dagneaux, E., S. Denness & S. Granger 1998. "Computer-aided Error Analysis", *System* 26, 163–74.

De Haan, P. 2000. "Tagging Non-native English with the TOSCA-ICLE Tagger", στο C. Mair & M. Hundt (επιμ.), *Corpus Linguistics and Linguistic Theory*. Άμστερνταμ: Rodopi, 69–79.

De Mönnink, I. 2000. "Parsing a Learner Corpus", στο C. Mair & M. Hundt (επιμ.), *Corpus Linguistics and Linguistic Theory*. Άμστερνταμ: Rodopi, 81–90.

Díaz-Negrillo, A. & J. Fernández-Domínguez. 2006. "Error Tagging Systems for Learner Corpora", *RESLA* 19, 83–102.

Dimitrakopoulou, M., G. Fotiadou, A. Roussou & I. M. Tsimpli. 2006. "Features and Agree Relations in L2 Greek", στο A. Belleti, E. Bennati, C. Chesi, E. Di Domenico & I. Ferrari (επιμ.), "*Language Acquisition and Development" Proceedings of GALA 2005*. Cambridge Scholars Press, 167–72.

—. 2007. "Features and Agree Relations in L2 Greek", *Proceedings of the 7th International Conference on Greek Linguistics* (ICGL7). University of York. [http://83.212.19.218/icgl7/Tsimpli-et-al.pdf]

Dimitrakopoulou, M., S. Kalaitzidou, A. Roussou & I. M. Tsimpli. 2004. "Clitics and Determiners in the L2 Greek Grammar", *Proceedings of the 6th International Conference on Greek Linguistics*. [http://www.philology.uoc.gr/conferences/6thICGL]

Dulay, H., M. Burt & S. Krashen. 1982. *Language Two*. Νέα Υόρκη: OUP.

Goutsos, D. 2010. "The Corpus of Greek Texts: A Reference Corpus of Modern Greek", *Corpora* 5(1), 29–44.

Granger, S. 1998. "The Computer Learner Corpus: A Versatile New Source of Data for SLA Research", στο S. Granger (επιμ.), *Learner English on Computer*. Λονδίνο: Addison Wesley Longman, 3–18.

—. 2002. "A Bird's Eye View of Learner: Corpus Research", στο S. Granger, J. Hung & S. Petch-Tyson (επιμ.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Άμστερνταμ: Benjamins, 3–33.

—. 2003. "The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research", *TESOL Quarterly* 37(3), 538–45.

Granger, S., E. Dagneaux & F. Meunier. 2002. *The International Corpus of Learner English*. Louvain-la-Neuve: Presses Universitaires de Louvain. [http://www.i6doc.com]

Hunston, S. 2002. *Corpora in Applied Linguistics*. Κέμπριτζ: Cambridge University Press.

Leech, G. 1992. "Corpora and Theories of Linguistic Performance", στο J. Startvik (επιμ.), *Directions in Corpus Linguistics*. Βερολίνο: Mouton de Gruyter, 105–22.

—. 1998. "Learner Corpora: What They Are and What Can Be Done with Them", στο S. Granger (επιμ.), *Learner English on Computer*. Λονδίνο: Addison Wesley, 14–20.

Meunier, F. 1998. "Computer Tools for Learner Corpora", στο S. Granger (επιμ.), *Learner English on Computer*. Λονδίνο: Longman, 19–37.

Milton, J. & N. Chowdhury. 1994. "Tagging the Interlanguage of Chinese Learners of English", στο L. Flowerdew & K. K. Tong (επιμ.), *Entering Text*. Hong Kong: The Hong Kong University of Science and Technology, 127–43.

Nicholls, D. 2003. "The Cambridge Learner Corpus: Error Coding and Analysis for Lexicography and ELT", *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, 572–81.

O'Donnell, M. 2008. "The UAM CorpusTool: Software for Corpus Annotation and Exploration", στο C. Bretones κ.ά. (επιμ.), *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, 1433–47.

O'Keeffe, A., M. McCarthy & R. Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Κέμπριτζ: Cambridge University Press.

Παπαδοπούλου, Δ. 2005. "Η παραγωγή της ρηματικής όψης στον προφορικό και το γραπτό λόγο σπουδαστών της ελληνικής ως δεύτερης/ξένης γλώσσας", *Journal of Applied Linguistics* 21, 39–54.

Pravec, N. A. 2002. "Survey of Learner Corpora", *ICAME Journal* 26, 81–114.

Selinker, L. 1972. "Interlanguage, IRAL", *International Review of Applied Linguistics in Language Teaching* 10(3), 209–31.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Οξφόρδη: Oxford University Press.

Tantos, A. & D. Papadopoulou. Υπό έκδ. "Stand-off Annotation in Learner Corpora. Compiling the Greek Learner Corpus (GLC)", στο A. D. Negrillo & F. J. D. Pérez (επιμ.), *Specialisation and Variation in Language Corpora*. Peter Lang International Academic Publishers.

Τζιμώκας, Δ. 2010. "Ηλεκτρονικό σώμα κειμένων (ΗΣΚ) εκμάθησης της νέας ελληνικής ως δεύτερης γλώσσας: προς ένα ερευνητικό και διδακτικό εργαλείο", *Proceedings of 30th Annual Meeting of the Department of Linguistics*. Θεσσαλονίκη: 602–16.

Tsimpli, I. M. 2003. "Clitics and Determiners in L2 Greek", στο J. M. Liceras, H. Zobl & H. Goodluck (επιμ.), *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference*. Somerville: Cascadilla, 331–39.

Tsimpli, I. M. & D. Papadopoulou. 2006. "Aspect and Argument Realisation. A Study on Antecedentless Null Objects in Greek", *Lingua* 116, 1595–615.

Tsimpli, I. M., A. Roussou, M. Dimitrakopoulou & S. Kalaitzidou. 2003. "The Production of Clitics and Articles by Slavic Speakers of Greek", *Proceedings of the 13th International Conference on Applied Linguistics*. Θεσσαλονίκη.

Van Rooy, B. & L. Schäfer. 2003. "Automatic POS Tagging of a Learner Corpus: The Influence of Learner Error on Tagger Accuracy", στο D. Archer, P. Rayson, A. Wilson & T. McEnery (επιμ.), *Proceedings of the Corpus Linguistics 2003 Conference* (CL 2003). Lancaster University, University Centre for Computer Corpus Research on Language, 835–44.

Λέξεις κλειδιά: Ελληνικό Σώμα Κειμένων Μαθητών (ΕΣΚΕΙΜΑΘ), σώματα κειμένων μαθητών, εξ αποστάσεως επισημείωση, πλαίσιο επισημείωσης λαθών.