The Corpus of Greek Aphasic Speech: Design and compilation

Dionysis Goutsos[1], Constantin Potagas[2], Dimitris Kasselimis[2 & 3],

Maria Varkanitsa[1] & Ioannis Evdokimidis[2]

[1]*Department of Linguistics, School of Philosophy, University of Athens*

[2]*Department of Neurology, Medical School, University of Athens, Aeginition Hospital*

[3]*Psychology Department, School of Social Sciences, University of Crete*

The paper presents the design and compilation of the Corpus of Greek Aphasic Speech, a new resource for the study of aphasia in Greek, and discusses its possible applications. The aims and design of the corpus and the methods followed for its compilation are presented. A pilot corpus, including two texts (spontaneous speech and picture description) from the spoken output of 20 patients was first created. On the basis of this, a classification of paraphasias or speech errors has been attempted and some preliminary findings have been gathered, while the target corpus is planned to include texts from 120 patients of about 50.000 words. It is argued that computer language corpora can offer a new perspective to the linguistic study of aphasia, especially in Greek, by providing systematic evidence for patients' linguistic profiles, drawn from language use rather than from isolated examples of lack of competence.

Keywords: annotation, aphasia, speech errors; Greek

En este artículo se presenta el diseño y la compilación del Corpus del Discurso Afásico Griego, un nuevo recurso para el estudio de la afasia en griego, y se analizan sus posibles aplicaciones. Los objetivos y el diseño del corpus y los métodos seguidos para su elaboración se presentan. Un corpus piloto, incluyendo dos textos (el habla espontánea y la descripción de imagen) del discurso de 20 pacientes fue creada. Sobre la base de esto, una clasificación de las parafasias o errores del habla se ha intentado y algunos resultados preliminares han sido recogidos, mientras que el corpus de destino está previsto incluir textos de 120 pacientes alrededor de 50.000 palabras. Se argumenta que un corpus puede ofrecer una nueva perspectiva para el estudio lingüístico de la afasia, especialmente en griego, proporcionando evidencia sistemática de los perfiles lingüísticos de los pacientes, basado en el uso del lenguaje en lugar de ejemplos aislados de falta de competencia.

Palabras clave: anotación, afasia, los errores del habla; griego

1. Corpora and aphasia

As late as in 1990 Menn & Obler remark that "there is absolutely no tradition of publication or even archiving of aphasic texts" (1990: 12). Although the collection of data from the speech of aphasic patients has had a long history in the study of aphasia, it is true that even today only a few studies of aphasia have taken full advantage of corpus methodology. Thus, most studies use databases of individual words or sentences rather than extended talk or use corpora in order to simply draw illustrative examples rather than as data to be exhaustively described. In addition, as Perkins has rightly pointed out, "most corpora of disordered language remain very small [...] hard to access [...] and – above all – not in machine-readable format" (1995: 129). Notable exceptions for adult aphasic discourse include Perkins & Varley (1996) for English, Gallardo & Moreno (2005) for Spanish and Westerhout & Monachesi (2007) for Dutch. Still, wide availability to data is only found in projects like the Aphasia TalkBank, which aims at the creation of an enormous multimedia database of aphasic speech in English and other languages (MacWhinney, 2007).

There are several reasons for this rather curious absence of corpus methodology from the study of aphasia. As Edwards finds, "most models applied to aphasia have been concerned with single-word processing" (2005: 23) rather than the analysis of extended talk. In fact, single-case studies predominate in the field, something which concurs with an almost exclusive focus on competence aspects of aphasic speech production and comprehension, tested through completion tasks, grammaticality judgements etc. In other words, the interest has been on what the aphasic patients are able (or, mainly, not able) to say rather than on what they have actually said in specific contexts. Without doubt, the predominance for several decades of competence-based models in linguistics, following the dominant generative research paradigm, has significantly weighted the scales against empirical investigations of aphasic speech performance.

This bias has particularly influenced the study of aphasia in Greek, which, as a highly inflected language, offers itself for the study of grammatical phenomena. This probably explains why most of the studies on Greek aphasia concern aspects of verb or noun morphology at the expense of the analysis of other linguistic levels. Moreover, these studies are exclusively based on competence testing rather than other evidence. Most importantly, there has been no attempt for an overall description of aphasia in Greek; instead, existing studies fall in the mainstream tradition of testing isolated phenomena, aiming at validating or disqualifying theoretical points.[16]

The advantages of using electronic corpora in the study of aphasia are quite obvious. First of all, corpora allow researchers to study large amounts of occurring data, by focusing on actual language use rather than linguistic competence. This is necessary if our goal is a comprehensive typology of aphasia based on linguistic principles (Crystal 2002). The collection of the spoken output of aphasic patients into a corpus enables us to treat linguistic deviations as instances of aphasic discourse, rather than as accidental, isolated

16 See Goutsos et al. (2011) for a detailed discussion of the literature on Greek.

errors and thus to account for their function in the patient's linguistic system. In addition, as both Perkins (1995) and Crystal (2002) emphasize, the occurrence of linguistic patterns may enable us to group patients into a number of linguistically defined diagnostic types, with a view to helping diagnosis, assessment and intervention. Furthermore, data from corpora of aphasic discourse can be used to compare with linguistic patterns in general reference corpora. In this way, we can assess the degree to which aphasic data diverge from other kinds of data on a firm empirical basis, rather than by relying on intuition. Finally, corpora enable easy accessibility to the original data and their transcription and/or annotation. They thus enhance verifiability of research, make possible cross-linguistic investigations and allow different kinds of researchers to study the same set of data. As hinted at above, there is an additional advantage from the use of aphasic corpora in the case of Greek, in which large-scale empirical findings on aphasia are still missing.

2. *The Corpus of Greek Aphasic Speech*

The Corpus of Greek Aphasic Speech (CGAS) is a development of a common project of the Department of Linguistics and the Department of Neurology (Aeginition Hospital) of the University of Athens. In the following sections we present the project's aims and describe the corpus composition and the stages of its compilation.

2.1. Corpus aims and composition

In designing the CGAS we have taken into consideration the relative absence of specialized corpora of aphasic discourse both in Greek and other languages, as well as the problems pointed out above the study of Greek aphasia. In addition, general principles of corpus design such as Sinclair's (2005) have been followed. In particular, the Corpus of Greek Aphasic Speech (CGAS) was compiled with the following aims:

a) to contain whole texts produced by a significant number of Greek aphasic patients,

b) to include texts from a variety of contexts,

c) to relate linguistic data to extra-linguistic metadata in a systematic way, and

d) to make data on Greek aphasic discourse available to the research community.

In particular, CGAS includes in its pilot phase data from 20 patients, who were treated at the Aeginition Hospital between 2006 and 2008. Two type texts from each patient's spoken output are included, namely spontaneous speech and picture description. In other words, the corpus includes 40 texts, two from each participant. Both text types were produced in the situation of doctor-patient interviews and, in this sense, they are not, strictly speaking, conversational, but rather guided monologues. In total, 12,663 words are included in the Corpus at present, of which 10,332 come from the patients' discourse.

Our plans for the development of CGAS aim at 120 patients, that is 240 texts from the same two text types. This will constitute a corpus of approximately 50,000 words, of which roughly 41,000 will come from the patients' speech. Table 1 below summarizes this information.

Table 1. Structure of the CGAS.

	Number of patients	*Text types*	*Number of texts*	*Word number*
Pilot corpus	20	2	40	12,663 (10,332)
Target corpus	120	2	240	~50.000 (41.000)

2.2. Corpus compilation

The stages of corpus compilation involved data and metadata collection, transcription and annotation, including marking for speech errors. The speakers were recruited from a large pool of patients with stroke treated at the Aeginition Hospital. Aphasics were identified through language assessment with the Boston Diagnostic Aphasia Examination–Short Form (BDAE-SF), adapted for Greek (Tsapkini et al. 2009). CT and/or MRI scans were obtained for each patient and lesion sites were identified by two independent neuro-radiologists.

The data were collected in typical doctor and patient interactions, in which psychologists interviewed patients with regard to what happened to them ('stroke stories') and, on another occasion, administered the description of the Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). Sessions were audio-recorded with either a tape-recorder or a digital voice recorder in a quiet setting. We have excluded data from patients who did not perform in either the spontaneous speech or the picture description task and from those who suffered severe fluency impairment, so that they could not produce any recognizable words.

All collected material was orthographically transcribed in a first transcript and then checked for accuracy by two different transcribers. (For reasons of transparency, an orthographic rather than phonetic transcription was adopted). However, fluency problems, voiced and unvoiced starters and fillers, repetitions and other phenomena of spoken interaction such as noise from the outside, coughing etc. were carefully noted, following conventions for spoken data transcription (Georgakopoulou & Goutsos, 2004: vii).

2.3. Corpus annotation

In the first phase of annotation, already transcribed data have been tagged for speech errors or paraphasias. These errors have also been tagged for part of speech. The second phase of annotation includes the extension of the part of speech tagging to the entire pilot

corpus, whereas at a later stage intonation contours will also be marked throughout. In parallel, the spoken data transcription is linked to the sound archives of the recordings, with a view to developing a multi-modal corpus that will allow flexibility in the analysis, as well as access to the representation of several layers of data.

Because of the gap in the comprehensive description of aphasia in Greek, it has been necessary to identify paraphasic errors at many levels of linguistic analysis, as has already been the case in the relevant bibliography (e.g. Nespoulous & Roch-Lecours, 1984; Ahlsén, 2006: 56-57; Ingram, 2007: 23; Turgeon & Macoir, 2008). Thus, the following categories and sub-categories of errors were identified: a) phonological errors, b) morphosyntactic errors, c) lexical errors, d) neologisms and e) periphrasis. Appendix 1 presents our detailed classification, along with the criteria used for each category and an indicative example from the data.

It must be noted here that there may be some overlap between the categories identified, due to the inevitable process of interpretation involved in the tagging for speech errors. For cases in which two categories could be attributed to the same error, it was decided to mark both, while a similar practice was followed in cases where two different errors were found to occur.

2.4. Corpus availability

At present, the annotated version of the pilot CGAS is available through the Aphasia Talkbank project.[17] For this release, data have been transcribed according to CLAN conventions (MacWhinney, 1996). Since this is a work in progress, we are currently using our experience from the pilot corpus to build the target corpus. Our plan is to make the target CGAS available to the research community via a dedicated webpage interface. Depending on ethical considerations (personal data restrictions), recordings will also be made available along with their transcription, so as to allow access to the primary material.

3. PRELIMINARY FINDINGS AND IMPLICATIONS

Our preliminary analysis of the pilot CGAS suggests that the corpus can be immensely helpful in the study of Greek aphasia. First of all, a new series of questions hitherto unexplored in the literature can now be raised with reference to authentic data of aphasic speech. In particular, information can be adduced on the frequency and types of phonological and lexical errors in Greek, including neologisms and other semantically-related errors. Linguistic phenomena such as periphrasis, which is indicative of the speaker's linguistic strategies, can now be placed alongside traditional types of paraphasias. In addition, the corpus can offer specific details about the particular errors occurring in Greek aphasic discourse. For instance, phonological omission errors seem to mainly concern consonant clusters in nouns, involving phonemes such as /r/, /s/, /a/ and /n/. These details are crucial,

17 For more details, see the project's site: http://talkbank.org/AphasiaBank

not only for a fully-fledged analysis of Greek aphasia, but also for orienting clinical and post-clinical intervention towards concrete findings.

Furthermore, a simple comparison of the ten most frequent words in the CGAS text types and related text types in a reference corpus of Greek such as the Corpus of Greek Texts (CGT, see Goutsos, 2010) can also be illuminating.

Table 2. Most frequent tokens in CGAS and CGT text types.

CGAS: interviews	CGAS: picture description	CGT: spoken data	CGT: interviews
και 'and'	*το* 'the'-NEUT	*και* 'and'	*και* 'and'
το 'the'-NEUT	*να* 'to'	*το* 'the'-NEUT	*να* 'to'
ε 'eh'	*είναι* 'is'	*να* 'to'	*το* 'the'-NEUT
να 'to'	*εδώ* 'here'	*ναι* 'yes'	*την* 'the'-FEM
μου 'my'	*ε* 'eh'	*είναι* 'is'	*είναι* 'is'
δεν 'not'	*αυτό* 'this'	*δεν* 'not'	*που* COMP
στο 'at'	*δεν* 'not'	*που* COMP	*ότι* COMP
με 'me'	*και* 'and'	*θα* 'will'	*η* 'the'-FEM
αυτό 'this'	*τα* 'the'-NEUT-PL	*τα* 'the'-NEUT-PL	*του* 'the'-GEN
ναι 'yes'	*τι* 'what'	*μου* 'my'	*της* 'the'-FEM

As can be seen in Table 2, aphasic data diverge from the reference corpus in interesting ways: hesitation ('eh') and vague words ('this', 'what') are more frequent, whereas the conjunction *και* ('and'), which is systematically the most frequent item in most text types in Greek, is very low in frequency in the picture description data. Furthermore, complementizers like *που* and *ότι* seem to be also much less frequent in CGAS.

A final remarkable aspect of aphasic speech concerns the use of word clusters or lexical bundles (Biber et al., 1999). In CGAS the most frequent clusters include phrases such as *δεν μπορώ/μπορούσα να το πω/να καταλάβω* 'I cannot/could not say/understand it', *πώς να το πω/τι να πω* 'how to say it/what can I say', *πρέπει να είναι* 'it must be', *αυτά εδώ πέρα/αυτό το πράγμα* 'these things/this thing over here'. These clusters are indicative of the discourse strategies followed by aphasic speakers (e.g. avoidance, modality, periphrasis) and can offer a first glimpse at formulaic language, which may be processed in different ways than the rest of the vocabulary in aphasia (Wray, 2002).

In conclusion, the development of the Corpus of Greek Aphasic Speech puts a much needed emphasis on spontaneously produced data and the analysis of speech errors in their discourse context. It thus allows assessing paraphasic errors as the product of situated language use by specific speakers rather than as isolated examples of lack of competence. Ease of access to both the original data and the levels of transcription, annotation etc., are also expected to significantly contribute to the understanding of aphasia in Greek and to enhance our knowledge of the field.

APPENDIX 1

The following categories of speech errors have been distinguished in the annotation of the pilot corpus:

1. Phonological paraphasias: errors affecting isolated phonemes or syllables.

   - PH1: phoneme deletion/omission: *άντας* ['adas], instead of *άντρας* ['adras] 'man'

   - PH2: phoneme addition: *αχαρτί* [aχa'rti], instead of *χαρτί* [χa'rti] 'paper'

   - PH3: phoneme substitution: *γρυκά* [γri'ka], instead of *γλυκά* [γli'ka] 'sweets'

   - PH4: syllabic: *σκαμπόβο* [ska'bovo], instead of *σκαμπό* [ska'bo] 'stool'

2. Morphosyntactic paraphasias: errors affecting grammatical morphemes.

   - MS1: morpheme deletion/omission: *αυτό άντρα* 'this man', instead of *αυτός είναι άντρας* 'this is a man'

   - MS2: morpheme addition: not found in the data

   - MS3: morpheme substitution: substitution (general): *για να πέσει κάτω*, instead of *θα πέσει κάτω* 'he will fall down' [the complementizer *για να* substitutes *θα*]

   - MS4: morpheme substitution: aspect: *δεν είδε καλά το μάτι*, instead of *δεν έβλεπε καλά το μάτι* 'the eye could not see' [synoptic/perfect in the place of continuous/ imperfect stem]

   - MS5: morpheme substitution: tense: *το λόγο που έχω πριν*, instead of *το λόγο που είχα πριν* 'the speech I had before' [present in the place of past stem]

   - MS6: morpheme substitution: agreement: *ένα κυρία* 'a.NEUT lady', instead of *μια κυρία* 'a.FEM lady'

   - MS7: other: *δουλεύω κάτι* 'I work something', instead of *δουλεύω σε κάτι* 'I work in something'

3. Lexical paraphasias: errors affecting whole words, particularly, substitution of a word by another pre-existing similar or non-similar word.

   - L1: formal: words related by formal similarity: *πλακάκι* [pla'kaci] 'tile', instead of *νεράκι* [ne'raci] 'some water'

   - L2: verbal: meaning similarity: *άνθρωπος* 'person', instead of *παιδί* 'child'

   - L3: unrelated: no similarity: *νούμερα* ['numera] 'numbers', instead of *μπισκότα* [bi'skota] 'biscuits'

4. Neologisms: errors affecting whole words (more than 50% of the word form): substitution of a word by another similar or non-similar word, not occurring in Greek.

   - N1: possible but non-existent words of Greek, classifiable to a part of speech: *γερεβύτης* [jere'vitis], instead of *νεροχύτης* [nero'çitis] 'basin'

   - N2: non-recognizable words, non-classifiable according to grammatical category: *πενιχθεσινίδις* [peniχθesin'iδis]

5. Periphrasis: errors affecting whole words: substitution of a word by an extended phrase.

   - P1: circumlocution: the extended phrase refers periphrastically to a word: *αυτό που έχει το νερό* 'this which has the water', instead of *βρύση* 'tap'

   - P2: vagueness: the extended phrase avoids specific reference to a word: *έπαθα μια αυτή* 'I had a this'

REFERENCES

AHLSÉN, E. (2006). *Introduction to Neurolinguistics*. Amsterdam/Philadelphia: John Benjamins.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S. & FINEGAN, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

CRYSTAL, D. (2002). CLINICAL LINGUISTICS. IN M. ARONOFF & J. REES-MILLER (Eds), *The Handbook of Linguistics* (pp. 673-682). Oxford: Blackwell.

EDWARDS, S. (2005). *Fluent Aphasia*. Cambridge: Cambridge University Press.

GALLARDO, B. & MORENO, V. (2005). *Afasia no fluente*. Valencia: Guada Impresores.

GEORGAKOPOULOU, A. & GOUTSOS, D. (2004). *Discourse Analysis. An Introduction*. (2nd ed.). Edinburgh: Edinburgh University Press.

GOODGLASS, H. & KAPLAN, E. (1983). *The Assessment of Aphasia and Related Disorders.* (2nd ed.). Philadelphia: Lea & Febiger.

GOUTSOS, D. 2010. The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora, 5* (1), 29-44.

GOUTSOS, D., POTAGAS, C., KASSELIMIS, D., VARKANITSA, M. & EVDOKIMIDIS, I. (2011). Studying paraphasias in the Corpus of Greek Aphasic Speech. In C. Potagas & I. Evdokimidis (Eds.), *Discourse and Memory* (pp. 23-47). Athens: Synapses. [In Greek]

INGRAM, J. C. L. (2007). *Neurolinguistics. An Introduction to Spoken Language Processing and its Disorders*. Cambridge: Cambridge University Press.

MACWHINNEY, B. (1996). The CHILDES system. *American Journal of Speech Language Pathology, 5*, 5-14.

MACWHINNEY, B. (2007). The TalkBank project. In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds.), *Creating and Digitizing Language Corpora* (pp. 163-180). London: Palgrave Macmillan.

MENN, L. & OBLER, L. K. (1990). Theoretical motivations for the cross-language study of agrammatism. In L. Menn & L. K. Obler (Eds.), *Agrammatic Aphasia: A Cross-language Narrative Sourcebook* (pp. 3-12). Amsterdam: John Benjamins. Vol. 1.

NESPOULOUS, J.-L. & ROCH LECOURS, A. (1984). Clinical descriptions of aphasia: Linguistic aspects. In D. Caplan, A. Roch Lecours & A. Smith (Eds.), *Biological Perspectives on Language* (pp. 141-157). Cambridge, Mass: MIT Press.

PERKINS, M. (1995). Corpora of disordered spoken language. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer. Transcription, Mark-up and Application* (pp. 128-134). London: Longman.

PERKINS, M. R. & VARLEY, R. (1996). *A Machine-Readable Corpus of Aphasic Discourse.* University of Sheffield: Department of Human Communication Sciences/Institute for Language, Speech and Hearing.

SINCLAIR, J. (2005). Corpus and text-basic principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1–16). Oxford: Oxbow Books.

TSAPKINI, K., VLAHOU, C. H. & POTAGAS, C. (2009). Adaptation and validation of standardized aphasia tests in different languages: Lessons from the Boston Diagnostic Aphasia Examination. *Behavioural Neurology, 22* (3), 111-119.

TURGEON, Y. & MACOIR, J. (2008). Classical and contemporary assessment of aphasia and acquired disorders of language. In B. Stemmer & H. A. Whitaker (Eds.), *Handbook of the Neuroscience of Language* (pp. 3-11). Amsterdam: Elsevier.

WESTERHOUT, E. & MONACHESI, P. (2007). A pilot study for a Corpus of Dutch Aphasic Speech (CoDAS). Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.1882&rep=rep1&type=pdf.

WRAY, A. (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

ACTAS DEL III CONGRESO INTERNACIONAL DE LINGÜÍSTICA DE CORPUS

LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES: PRESENTE Y FUTURO EN EL ANÁLISIS DE CORPUS

Editores:
María Luisa Carrió Pastor
Miguel Ángel Candel Mora

Editores
María Luisa Carrió Pastor
Miguel Ángel Candel Mora

ACTAS DEL III CONGRESO INTERNACIONAL DE LINGÜÍSTICA DE CORPUS.

LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES: PRESENTE Y FUTURO EN EL ANÁLISIS DE CORPUS

EDITORIAL
UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Esta editorial es miembro de la UNE, lo que garantiza la difusión y comercialización de sus publicaciones a nivel nacional e internacional.

Primera edición, 2011

ISBN: 978-84-694-6225-6
Ref. editorial: 6032