DOI:10.30842/ielcp230690152203

Т. А. Архангельский, М. Л. Кисилиер

КОРПУСА ГРЕЧЕСКОГО ЯЗЫКА: ДОСТИЖЕНИЯ, ЦЕЛИ И ЗАДАЧИ*

Принято считать, что эллинистика относится к наиболее консервативным областям гуманитарного знания. Тем не менее, одним из лучших на данный момент лингвистических корпусов заслуженно считается корпус древнегреческого языка, известный как Thesaurus Linguae Graecae. В то же время для новогреческого одновременно разрабатывается несколько корпусов, сильно отличающихся друг от друга по размерам, принципам работы и целям. В настоящей статье предпринимается попытка сопоставить разные корпуса и определить перспективы развития Корпуса греческого языка, разработанного на базе ИЛИ РАН и Греческого института СПбГУ.

Ключевые слова: корпусная лингвистика, Thesaurus Linguae Graecae, корпуса новогреческого языка, греческая диглоссия.

1. Вводные замечания

Так сложилось, что научные дисциплины, которые можно объединить под общим названием «эллинистика», на фоне других гуманитарных дисциплин отличаются заметной консервативностью. Попытки использования современных методов исследования часто подвергаются безжалостной критике, и во-многом эта критика, действительно, справедлива. Тем более удивительным кажется тот факт, что именно в эллинистике, и ранее всего в классической филологии, стали использоваться корпусные методы исследования[1].

Корпусная лингвистика – относительно молодая область. Первые корпуса появились только 1960-е гг., например, корпус английского языка (Kučera, Nelson 1967). Их отличали небольшой размер и отсутствие технических средств поиска, поэтому они скорее могли использоваться для получения разнообразных статистических данных, нежели помочь при проведении линг-

* Исследование выполнено при поддержке гранта РФФИ № 18-012-00607 «Корпусные и полевые исследования сентенциальных дополнений в балканских языках и диалектах».

[1] Впрочем, вероятно, создание электронного корпуса как раз и соответствовало традициям классической филологии. Еще в 1572–1573 гг. Анри II Этьен (Стефанус) создал лексикон, известный как «Thesaurus Linguae Graecae». Подобные труды создавались и в XVIII в.

вистического/филологического анализа. Корпуса современного типа появились около 1990-х гг. Обычно величина «новых» корпусов варьирует между 1 млн. и 1 млрд. словоупотреблений[2], хотя встречаются и исключения.

В настоящей статье речь пойдет о корпусах, связанных с греческим языком. При обзоре существующих корпусов впервые будет представлено описание Корпуса греческого языка, разработанного ИЛИ РАН совместно с Греческим институтом СПбГУ.

2. Thesaurus Linguae Graecae (TLG)

Одним из наиболее успешных корпусов современного типа является знаменитый корпус древнегреческого языка TLG[3], созданный в университете Калифорнии (Berkowitz, Johnson 1990). Проект по подборке электронной библиотеки для корпуса начался в 1972 г.; первые версии TLG появились на кассетах в 1976 г., а компакт-диски вышли в 1985 (27 млн. слов)[4], 1988 (42 млн. слов), 1992 и 2000 гг. После успешного выхода второго диска в 1988 г. консультативный совет под председательством И. И. Шевченко одобрил расширение корпуса за счет схолий, а также византийской историографии и лексикографии. В 2001 г. стало ясно, что необходимо переходить от компакт-дисков к онлайн-версии, что открыло перед разработчиками TLG абсолютно новые перспективы. С одной стороны, перевод корпуса в интернет (2004 г.) позволил довести его до XIV в., а с другой стороны, дал возможность ввести с 2006 г. поиск по леммам. Сейчас TLG включает более 105 млн. словоупотреблений из более чем 12000 произведений и наглядно демонстрирует, что отличает корпус от электронной библиотеки – возможность настройки разнообразных параметров поиска, в том числе и синтактико-грамматических. К важнейшим особенностям TLG, помимо поиска лемм и словоформ с учетом метаданных текстов, таких как имя автора, название текста и хронологические рамки, можно отнести выход на словари древнегреческого (например, Liddell, Scott 1996) и новогреческого языков (например, Τριανταφυλλίδης

[2] Например, объем Национального корпуса русского языка (http://ruscorpora.ru/) превышает 600 млн. словоупотреблений.

[3] Полное название — Thesaurus Linguae Graecae: A Digital Library of Greek Literature, http://stephanus.tlg.uci.edu/.

[4] Этот диск стал первым в истории немузыкальным компакт-диском (http://stephanus.tlg.uci.edu/history.php, дата обращения: 07.05.18).

2009), возможность просмотреть все формы слова, представленные в корпусе, статистику употребления по векам и по авторам, а также географическую дистрибуцию.

Несмотря на очевидные достоинства, у TLG есть существенный недостаток – в нем использованы далеко не лучшие издания и не учитывается критический аппарат, крайне важный для исследований по классической филологии и по византинистике. Это не позволяет рассматривать TLG в качестве полноценного филологического корпуса (ср. Казанский 2016). С одной стороны, описанная проблема легко преодолима с помощью обращения к новейшим изданиям, по которым сверяются примеры, обнаруженные благодаря поиску в корпусе. Подобный подход, требуя некоторых усилий (поиск соответствующего издания, чаще всего недоступного онлайн), обладает, по крайней мере, одним несомненным достоинством — он заставляет исследователя обращаться к тексту, а не довольствоваться исключительно примером, предоставленным корпусом. С другой стороны, уже существуют положительные примеры того, как в корпусе могут быть учтены комментарии, схолии, рукописные варианты и конъектуры[5], что значительно повышает полезность корпуса, позволяя вести поиск не только по основному тексту произведения. К сожалению, подобная возможность едва ли реализуема в рамках TLG, который, будучи дорогим коммерческим проектом, не может претендовать на бесплатное использование чужих материалов. Важная особенность TLG – доступ по платной подписке.

3. Корпус греческих текстов (ΣΕΚ) и Корпус разговорного новогреческого языка (СПΛ)

Сейчас TLG является фактически единственным корпусом древнегреческого и средневекового греческого[6]. Для новогреческого же языка ситуация обстоит принципиально иначе. Нам известны, по крайней мере, пять корпусов, правда, не все из них функционируют на данный момент. О Корпусе греческих

[5] http://www.catullusonline.org/CatullusOnline/index.php (дата обращения: 08.05.18). Авторы статьи благодарят Н. Н. Казанского, познакомившего их с данным ресурсом.

[6] Мы сознательно избегаем здесь термина ‘среднегреческий язык’ в связи с высокой степенью языковой вариативности средневековых текстов, одни из которых очевидным образом ориентированы на древнегреческую традицию, а другие, напротив, очень похожи на новогреческое состояние (Eideneier 1972; 2005a; 2005b; Καπλάνης 2002).

текстов (Σώμα Ελληνικών Κειμένων)[7] известно из подробного описания Дионисиса Гуцоса (Goutsos 2010). Согласно этому описанию, Корпус греческих текстов достигает 30 млн. словоупотреблений; в нем представлены тексты разных жанров (в том числе значительное место занимают разговорные и устные тексты), а также тексты на кипрском диалекте.

Намного сложнее судить о Корпусе разговорного новогреческого языка (Corpus Προφορικού Λόγου)[8]. Согласно описанию, представленному на сайте корпуса, он был разработан в Институте новогреческих исследований Университета им. Аристотеля (Салоники, Греция) в рамках проекта, направленного на изучение речевой деятельности и диалогов с точки зрения дискурсивного анализа (рук. проекта Ф.-С. Павлиду). Значительную часть этого ресурса составляют оцифрованные аудио- и видеоматериалы (ок. 190000 Мб), а размер их дешифровок доходит до 1,9 млн. словоупотреблений. Материал для корпуса собирался в процессе записи живых диалогов между друзьями и родственниками, на учебных занятиях, во время разговоров по телефону, а также теле- и радиопередач.

4. Национальный тезарус греческого языка (ΕΘΕΓ)

Одним из старейших функционирующих корпусов является Национальный тезаурус греческого языка (Εθνικός Θησαυρός Ελληνικής Γλώσσας)[9], созданный в Институте обработки слова (Ινστιτούτο Επεξεργασίας του Λόγου) в Афинах. Сейчас он включает в себя ок. 47 млн. словоупотреблений и является корпусом исключительно письменных текстов. Все тексты, включенные в корпус, были написаны не ранее 1990 г. По большей части, это газеты и книги, имеющие широкий круг читателей. При создании поискового запроса можно учитывать метатекстовые категории, ориентирующиеся на жанр/тип текстов и на их тематику/содержание. В Национальном тезаурусе греческого языка предусмотрены несколько разновидностей поиска: по словоформам, по леммам и по довольно ограниченным грамматическим характеристикам, учитывающим частеречную принадлежность и некоторые дополнительные характеристики. Разные варианты поиска можно комбини-

[7] http://www.sek.edu.gr/index.php?el.
[8] http://ins.web.auth.gr/index.php?option=com_content&view=article&id=506:corpus-speech159&catid=40&lang =el&Itemid=165.
[9] http://hnc.ilsp.gr (дата обращения: 08.05.2018).

ровать. При поиске словосочетания лексемы, входящие в его состав, должны вводиться не в одно поисковое окно, а в разные. Корпус не позволяет искать словосочетания, состоящие более чем из трех слов. Также нельзя задавать расстояние между компонентами словосочетаний. Доступ в Национальный тезаурус греческого языка платный, однако существует бесплатная версия с ограниченным функционалом.

5. Греческий интернет-корпус (elTenTen)

Очень интересным проектом представляется Греческий интернет-корпус (The Greek Web Corpus, elTenTen)[10]. Он построен на текстах, собранных в интернете, и поражает своим объемом – 1,6 млрд. слов на август 2014 г. Греческий интернет-корпус открывает перед пользователем очень широкие возможности поиска: по словоформам, по леммам, по словосочетаниям, по синонимам и проч. При этом могут быть учтены и определенные грамматические характеристики. К сожалению, в корпусе практически отсутствует метатекстовая информация. При поисковом запросе нельзя установить ограничения по типу/жанру текста или указать хронологические рамки.

6. Корпус греческого языка (КГЯ)

В 2011–2015 гг. при поддержке Программы фундаментальных исследований Президиума Российской академии наук «Корпусная лингвистика» возникли 14 корпусов[11], и среди них Албанский национальный корпус[12] и Корпус греческого языка[13]. Все эти корпуса работают на платформе, созданной для Восточноармянского национального корпуса (ВАНК)[14], и принципиально отличаются от прочих корпусов тем, что их целевой аудиторией являются люди, незнакомые с корпусной лингвистикой и решающие как сложные лингвистические, так и сугубо

[10] https://www.sketchengine.co.uk/eltenten-greek-corpus/ (дата обращения: 08.05.18).
[11] Большинство из них сейчас находятся по адресу web-corpora.net (дата обращения: 08.05.18).
[12] http://web-corpora.net/AlbanianCorpus/search/?interface_language=ru (дата обращения: 08.05.2018; подробнее см. Morozova, Rusakov 2014; Морозова и др. 2016).
[13] http://web-corpora.net/GreekCorpus/search/?interface_language=ru (дата обращения: 08.05.18).
[14] http://www.eanc.net/EANC/search/?interface_language=ru (дата обращения: 08.05.18)

практические задачи. Это стало возможным благодаря очень простому интерфейсу. Другая общая черта всех корпусов, созданных при поддержке Президиума РАН, – это свободный доступ для всех пользователей. Во избежание нарушения авторских прав пользователь не может просматривать тексты целиком; ему открыт контекст до трех предложений перед и после искомого слова.

Корпус греческого языка на данный момент составляет 37,5 млн. словоупотреблений. Коллекция используемых текстов состоит по большей части из газет, но также имеются произведения почти 50 греческих и иностранных авторов XIX–XX вв. В отдельном подкорпусе собрана переводная литература. Возможен поиск по отдельным авторам или даже произведениям. В корпусе предусмотрены довольно широкие параметры поиска: кроме поиска по словоформам и леммам, можно искать по грамматическим формам и даже по английскому переводу. Крайне полезной представляется опция поиска сочетаний слов с указанием любого расстояния между ними. При этом количество слов, входящих в словосочетание, неограничено.

Одной из важнейших проблем, связанных с историей греческого языка, является знаменитый греческий вопрос, частично отразившийся в греческой диглоссии. В Корпусе греческого языка впервые была предпринята попытка учесть проблематику греческого языкового вопроса. Так, пользователь может уточнить интересующий его языковой вариант (кафаревуса или димотика) и орфографию (монотоническая или политоническая).

7. Заключение

Очевидно, что ни один из рассмотренных выше корпусов не может пока претендовать на то, чтобы стать единственным/основным корпусом греческого языка. У каждого из них есть свои особенности, связанные с охватом материала и размерами, а также свои технические особенности, сопоставление которых представлено в Таблице 1

Таблица 1. Сопоставление корпусов

	TLG	**ΣΕΚ**	**ΣΠΛ**	**ΕΘΕΓ**	**elTenTen**	**ΚΓΗ**
Свободный доступ	–	+?	+?	–	–	+
Языки интерфейса	1	1?	2	1	1	3
Поиск по леммам	+	–?	–?	+	+	+
Грамматический поиск	–	+/–?	–?	+/–	+/–	+
Просмотр парадигмы	+	–?	–?	–	–	–
Поиск синонимов	–	–?	–?	–	+	–
Статистика	+	–?	–?	+	+	–
Конкорданс	+	–?	–?	+	+	–
Связь со словарем	+	–?	–?	–	–	–
Перевод лексем	–	–?	–?	–	–	+
География/диалекты	+	+?	–?	–	–	–
Метаданные	+	+?	+?	+/–	–	+
Расстояние между лексемами при поиске	–	–?	–?	–	–	+
Выбор языкового варианта	–	–?	–?	–	–	+
Выбор типа/жанра текста	–	–?	–?	–	–	+
Поиск по отдельным авторам	+	–?	–?	–	–	+
Поиск по отдельным текстам	+	–?	–?	–	–	+
Подкорпус переводных текстов	–	–?	–?	–	–	+
Грамматический разбор	–	–?	–?	–	–	+

Даже это весьма поверхностное сопоставление демонстрирует, что несомненным образцом для греческих корпусов (при всех своих минусах) является Thesaurus Linguae Graecae. Новогреческие же корпуса при всем своем многообразии практически не пересекаются друг с другом. Поскольку все они построены на разных платформах, а многие из них являются коммерческими проектами, их объединение в один корпус едва ли представляется возможным. Уместно предположить два варианта развития событий в будущем:

1. Все корпуса продолжат развиваться параллельно, и пользователь должен будет учитывать их все. Ничего особенно страшного в этом нет, однако, во-первых, будет невозможно получить надежную статистику, а во-вторых, часть текстов окажется вне сферы внимания существующих корпусов, что вполне вероятно приведет к увеличению количества корпусов новогреческого языка. Похожим образом складывается ситуация в новогреческой лексикографии, где существует несколько больших словарей (например, Μπαμπινιώτης 2002; Τριανταφυλλίδης 2009), однако остается широкий пласт лексики, в том числе из ключевых литературных произведений, не учтенный ни в одном из существующих словарей (ср. Mackridge 2002).

2. Один из корпусов постепенно включит в себя материалы, используемые остальными корпусами. Едва ли такими корпусами могут стать Корпус разговорного новогреческого языка (СПΛ), Национальный тезаурус греческого языка (ΕΘΕΓ) и Греческий интернет-корпус (elTenTen), поскольку у первого другие задачи, второй направлен исключительно на новые тексты и в нем не предусмотрены инструменты для решения многих проблем, связанных с анализом новогреческих текстов XIX–XX вв. Последний же, будучи интернет-корпусом, способен обрабатывать лишь тексты, выложенные в свободном доступе. Не имея доступа к Корпусу греческих текстов (ΣΕΚ), мы не станем рассуждать о перспективах его развития, и остановимся здесь лишь на Корпусе греческого языка.

Даже в современном своем состоянии, несмотря на многие серьезные недостатки, он может включить тексты любого типа/жанра, а при подготовке соответствующих грамматик его хронологические рамки могут быть значительно расширены. В 2017 г. была начата разработка новой платформы для корпуса. На данный момент новая платформа завершена и проходит тестирование. По окончании тестирования на нее будут переведены ряд корпусов, включая Корпус греческого языка и Албанский национальный корпус. Предполагается также, что на базе этой платформы будут созданы корпуса цаконского диалекта и диалекта приазовских греков, в которые наравне с письменными текстами войдут и аудиоматериалы. Ожидается, что в этих корпусах будет также возможен поиск по лексико-семантическим группам и выход на словари. Новая платформа, помимо повышенной производительности, позволяет получать разнообразные статистические данные, парадигмы словоформ, встречающихся в корпусе, с учетом их частотности. В ней будут применены новые методы для решения греческого языкового вопроса. Однако самым важным для успешного развития корпуса является увеличение его объема за счет включения новых текстов. Конечно, он никогда не сможет достигнуть размеров Греческого интернет-корпуса (elTenTen), однако для получения более или менее представительных результатов он должен включать как можно больше текстов.

Литература

Babiniotis, G. 2002: *Lexiko tis Neas Ellinikis glossas me skholia gia ti sosti skhesi ton lexeon* [*Dictionary of Modern Greek with comments on the correct word usage*]. 2nd ed. Athens: Centre of lingistics.

Μπαμπινιώτης, Γ. 2002: *Λεξικό της Νέας Ελληνικής γλώσσας με σχόλια για τη σωστή σχέση των λέξεων*. 2 εκδ. Αθήνα: Κέντρο γλωσσολογίας Ε.Π.Ε.

Berkowitz, L., Johnson, W. H. 1990: *Thesaurus Linguae Graecae Canon of Greek authors and works*. 3rd ed. New York: OUP.

Eideneier, H. 1972: [Thoughts on monotonic system]. *Mandatophoros* 14, 16–19.

Eideneier, H. 1972: Σκέψεις σχετικά με το μονοτονικό σύστημα. *Μαντατοφόρος* 14, 16–19.

Eideneier, H. 2005a: [Orthographic anarchy — lack of education? Orthographic problems in Post-Byzantine manuscripts]. *Studies in Greek linguistics* 25, 197–205.

Eideneier, H. 2005a: Ορθογραφική αναρχία — έλλειψη παιδείας; Ζητήματα ορθογραφίας σε μεταβυζαντινά χειρόγραφα. *Μελέτες για την Ελληνική γλώσσα* 25, 197–205.

Eideneier, H. 2005b: [Orthophony vs. Orthography]. In: Jeffreys, E. M., Jeffreys, M. J. (eds.). *Approaches to texts in early Modern Greek*. Oxford: Oxford University Press, 3–16. (Neograeca Medii Aevi. Anadromika kai Prodromika. Vol. 5).

Eideneier, H. 2005b: Ορθοφωνία vs. Ορθογραφία. In: Jeffreys, E. M., Jeffreys, M. J. (eds.). *Approaches to texts in early Modern Greek*. Oxford: Oxford University Press, 3–16. (Neograeca Medii Aevi. Αναδρομικά και Προδρομικά. Τ. 5).

Goutsos, D. 2010: The corpus of Greek texts: a reference corpus for Modern Greek. *Corpora* 5. 1, 29–44.

Kaplanis, T. A. 2002: [Editions of the texts of Modern Greek literature in vernacular and monotonic system]. *Kondylophoros* 1, 205–235.

Καπλάνης, Τ. Α. 2002: Εκδόσεις κειμένων της νεοελληνικής δημώδους γραμματείας και μονοτονικό σύστημα. *Κοντυλοφόρος* 1, 205–235.

Kazansky, N. N. 2016: [Problems of creation of philological corpus]. In: Tishkov V. A. (ed.). *Trudy Otdeleniya istoriko-filologicheskih nauk. 2015*. Moscow: Nauka, 133–146.

Казанский, Н. Н. 2016: Проблемы создания филологического корпуса. В сб.: Тишков, В. А. (ред.). *Труды Отделения историко-филологических наук. 2015*. М.: Наука, 133–146.

Kučera, H., Nelson, F. W. 1967: *Computational analysis of present-day American English*. Providence: RI: Brown University press.

Liddell, H. G., Scott, R. 1996. *A Greek-English lexicon*. Oxford: Clarendon Press.

Mackridge, P. H. 2002: [G. Babiniotis, Dictionary of Modern Greek. Athens: Kentro Lexikologias, 1998. 2064 pages. Dictionary of Common Modern Greek. Thessaloniki: Aristoteleio Panepistimio Thessalonikis, Institouto Neoellinikon Spoudon [Idryma Manoli Triandafyllidi], 1998. xxxii + 1532 pages]. *Journal of Greek Linguistics* 2. 1, 254–259.

Mackridge, P. H. 2002: G. Babiniotis, Λεξικό της νέας ελληνικής γλώσσας. Athens: Kentro Lexikologias, 1998. 2064 pages. Λεξικό της κοινής νεοελληνικής. Thessaloniki: Aristoteleio Panepistimio Thessalonikis, Institouto Neoellinikon Spoudon [Idryma Manoli Triandafyllidi], 1998. xxxii + 1532 pages. *Journal of Greek Linguistics* 2. 1, 254–259.

Morozova, M. S., Arkhangelskiy, T. A., Daniel, M. A., Rusakov, A. Yu. 2016. [Albanian National Corpus: main lines of work]. *Acta linguistica Petropolitana* XII. 3, 171–191.

Морозова, М. С., Архангельский, Т. А., Даниэль, М. А., Русаков, А. Ю. 2016: Албанский национальный корпус: основные направления работы. *Acta linguistica Petropolitana. Труды Института лингвистических исследований РАН* XII. 3, 171–191.

Morozova, M., Rusakov, A. 2015: Albanian National Corpus: composition, text processing and corpus-oriented grammar development. In: Demiraj, B. (hrsg.). *Sprache und Kultur der Albaner. Zeitliche und räumliche Dimensionen. Akten der 5. Deutsch-albanischen kulturwissenschaftlichen Tagung (5.–8. Juni 2014, Buçimas bei Pogradec, Albanien)*. Wiesbaden: Harrassowitz Verlag, 270–308..

Triantafyllidis, M. 2009: *Lexiko tis Koinis Neoellinikis* [*Dictionary of Common Modern Greek*]. 8th ed. Aristotle University of Thessaloniki: Institute of Modern Greek Studies.

Τριανταφυλλίδης, Μ. 2009: *Λεξικό της Κοινής Νεοελληνικής*. 8 εκδ. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης; Ινστιτούτο Νεοελληνικών Σπουδών.

T. A. Arkhangelskiy, M. L. Kisilier. Corpora of Modern Greek: achievements and goals

Methods of corpus linguistics become more and more important in Modern Greek Studies. More than forty years ago appeared first versions of Thesaurus Linguae Graecae that has now become one of the most elaborated and functional corpora in the world despite some drawbacks. There are at least five corpora that are relevant for Modern Greek. Most of them have different types of data, different size (from 1.9 million tokens up to 1,6 billion) and are designed for different tasks. The comparison of these corpora demonstrates that none of them can now fully replace the others, however it is not likely that all these corpora may be developed simultaneously. In this article we tried to describe the Corpus of Modern Greek (http://web-corpora.net/GreekCorpus/search/?interface_language=en) and its unique features in order to demonstrate why and how it could undertake functions of the most corpora of Modern Greek, except, probably, the Greek Web Corpus, elTenTen. Unlike other Greek corpora, the Corpus of Modern Greek was created by linguists for linguists and for non-professional users and does not require any special registration. Its structure allows it to work with different types of texts including audio data. It possesses a powerful search engine which enables to take into account many detailed grammatical features. Apart from that, the user of the Corpus of Modern Greek can find here translations from English into Modern Greek. We hope that in the nearest future this option will be relevant for Russian as well.

Keywords: corpus linguistics, Thesaurus Linguae Graecae, Modern Greek Corpora, Corpus of Modern Greek, Greek diglossia.