Lang Resources & Evaluation
DOI 10.1007/s10579-010-9130-z

ORIGINAL PAPER

IPLR: an online resource for Greek word-level and sublexical information

Athanassios Protopapas · Marina Tzakosta · Aimilios Chalamandaris · Pirros Tsiakoulis

Abstract We present a new online psycholinguistic resource for Greek based on analyses of written corpora combined with text processing technologies developed at the Institute for Language & Speech Processing (ILSP), Greece. The "ILSP PsychoLinguistic Resource" (IPLR) is a freely accessible service via a dedicated web page, at http://speech.ilsp.gr/iplr. IPLR provides analyses of user-submitted letter strings (words and nonwords) as well as frequency tables for important units and conditions such as syllables, bigrams, and neighbors, calculated over two word lists based on printed text corpora and their phonetic transcription. Online tools allow retrieval of words matching user-specified orthographic or phonetic patterns. All results and processing code (in the Python programming language) are freely available for noncommercial educational or research use.

Keywords Online resources · Text corpora · Sublexical variables · Psycholinguistics · Greek · Syllabification · Bigrams

We present a new online psycholinguistic resource for Greek based on analyses of written corpora combined with text processing technologies developed at the Institute for Language & Speech Processing (ILSP), Greece. The "ILSP Psycho-Linguistic Resource" (IPLR) is a freely accessible service via a dedicated web page, at http://speech.ilsp.gr/iplr, providing analyses of user-submitted letter strings (words and nonwords) as well as frequency tables for important units and sets such as syllables, bigrams, and neighbors. The purpose of this document is to announce

A. Protopapas (✉) · A. Chalamandaris · P. Tsiakoulis
Institute for Language and Speech Processing, Artemidos 6 & Epidavrou, 15125 Maroussi, Greece
e-mail: protopap@ilsp.gr

M. Tzakosta
University of Crete, Rethymno, Crete, Greece

Published online: 01 September 2010

and introduce this resource to the research community, in the context of recent related work in psycholinguistic resources. This article serves as reference for IPLR in general and presents some basic information on descriptive statistics and frequency measures. We avoid repetition of detailed information and results available from the web site.

1 Resources for psycholinguistic experimentation

Psycholinguistic experimentation requires metrics and other quantitative information at lexical and sublexical levels, on the basis of which to select or match stimuli, as well as to use as predictors for various performance metrics. The availability of large amounts of electronic text and the accessibility of programming and internet technologies allow calculation of many important properties and construction of interactive interfaces for remote use.

To serve these needs, a number of online corpora and tools are now available for several languages. For example, WebCelex (http://celex.mpi.nl/) is an online web interface for the CELEX lexical databases of the Max Planck Institute for Psycholinguistics including English, Dutch, and German databases with associated orthographic, phonological, morphological, syntactic, and frequency information (Baayen et al. 1995). WebCelex allows complex searches over the database contents at the level of lemmas, word forms, and syllables. The English Lexicon Project (ELP) at Washington University in St. Louis provides lexical characteristics and behavioral data from visual lexical decision and naming studies of more than 40,000 words (http://elexicon.wustl.edu/; see Balota et al. 2007). The ELP web site provides lists of words and nonwords matching criteria specified online by the user. These resources facilitate selection of stimuli and control of their properties.

Some resources are available for other languages as well. For example, in Italian, LEXVAR (Barca et al. 2002) provides psycholinguistic norms for 626 "simple nouns" and CFVlexvar (Rinaldi et al. 2004) lists properties of the first words acquired by children, while CoLFIS constitutes a tagged representative corpus of 3.8 million tokens from printed texts with associated frequency counts and online search facility (http://www.istc.cnr.it/material/database/colfis/index_eng.shtml, described in Laudanna et al. 1995). In Portuguese, Porlex provides orthographic, phonological, phonetic, part-of-speech, and neighborhood information for about 30,000 words (Gomes and Castro 2003). Several online databases are available for French, including Novlex (http://www2.mshs.univ-poitiers.fr/novlex/; Lambert and Chesnet 2001) and Manulex (http://leadserv.u-bourgogne.fr/bases/manulex/manulexbase/; Lété et al. 2004), based on grade-level elementary school book corpora, and Lexique (http://www.lexique.org/; New et al. 2004), based on a corpus of contemporary literary texts. Manulex_Infra (http://leadserv.u-bourgogne.fr/bases/manulex/manulex_infra/; Peereman et al. 2007) adds to the Manulex corpus sublexical variable information of interest to psycholinguistics researchers. Similar to ELP, Brulex and Lexique provide online search interfaces to allow retrieval of selected word forms based on specific properties set by user-specified constraints.

In Greek, existing resources include the ILSP Hellenic National Corpus (HNC, see http://hnc.ilsp.gr/; Hatzigeorgiu et al. 2000) and GreekLex (Ktori et al. 2008) at the University of Nottingham (http://www.psychology.nottingham.ac.uk/GreekLex/). The HNC constitutes a major effort in collecting and annotating printed texts from multiple modern sources, and allows online searches for multiple words, providing phrase context and concordance information. The HNC includes a morphological analysis tool mapping word forms to lemmas, allowing word families to be retrieved simultaneously and treated uniformly. The HNC is used for many types of linguistic analysis and serves technology development in the domains of electronic lexicography and machine translation. However, the HNC does not provide intramorphemic analyses and no counts other than lemma and token frequencies of occurrence (under "Statistics" at http://hnc.ilsp.gr/en/statistics.asp). A recent resource for research concerning early stages of learning to read, the Educational Hellenic Corpus (http://www.xanthi.ilsp.gr/ethek/) provides information similar to the HNC, based on a corpus of current elementary school books.

Reading research often requires consideration of the frequencies of sublexical units, such as individual letters and bigrams (i.e., pairs of adjacent letters). GreekLex is a first attempt at providing neighborhood and bigram information for Greek words, based on the conjunction of HNC base forms (lemmas) and an online dictionary (resulting in a total of 35,304 word forms). Neither the HNC nor GreekLex contain pronunciation information, therefore up to now it has not been possible to consider phonological neighbors, bigrams, or syllables in Greek. Research on spoken language has thus been lacking a quantitative basis for the relevant lexical and sublexical units and properties affecting speech production and perception.

IPLR is a recent effort based on a text-to-speech approach to word forms found in printed text corpora, thus permitting manipulations and computations on the phonetic domain. It was designed to address the specific lack of Greek resources in this domain and to complement pre-existing approaches by adding more comprehensive sets of measures, both orthographic and phonetic. IPLR provides information for an unrestricted set of lexical forms including all inflected variants, not only lemma base forms, and provides an online search and retrieval interface for lexical and sublexical properties, as already available for corpora in other languages.

2 Sources and analyses of IPLR

IPLR provides data and services based on two printed text corpora (previously presented in Protopapas 2006), including a very large corpus (272 million tokens) made up entirely of journalistic texts ("L corpus"), and a smaller corpus (34 million tokens) including journalistic, legal, and literary texts from HNC ("C corpus"). The latter has been checked to some extent, and verified against an online Greek dictionary, whereas the former contains raw texts including numerous typographical and other errors. Both corpora have been pre-processed to remove letter strings including numerals, symbols, or non-Greek letters. We recommend using the C corpus (denoted "Clean" in the online tool pages) for all tasks not specifically

requiring inclusion of proper names and foreign words. When using the L corpus (denoted "Large"), keep in mind that a large proportion of retrieved types are likely misspelled or otherwise incorrect.

All context information from the printed texts has been discarded, therefore both corpora are essentially lists of individual word forms (types), with associated (token) counts, as far as the IPLR is concerned. The orthographic word forms have been converted to phonetic forms with a module developed for Greek text-to-speech synthesis, which is used in many commercial applications and is known to perform very well (99.4% word-level success rate with out-of-vocabulary words, and 98.5% with untrained proper names; Chalamandaris et al. 2005). Phonetically ambiguous orthographic forms (which can be pronounced in more than one way) have been checked manually and corrected when necessary (see Protopapas and Vlahou 2009, for explanation and quantification of graphophonemic ambiguities in the Greek orthography). Thus, a set of pairs was constructed, each consisting of the orthographic and phonetic representation of a single word type. Because the basis of this set was the printed text, and only one pronunciation was derived for each item, homophones (spelled differently but pronounced identically) are included; however, homographs (spelled identically but pronounced differently) are impossible to detect and therefore essentially discarded.

The level of segmental analysis employed in IPLR is the broad phonetic level of "speech sound" categories, as they occur in canonical pronunciation typical of major cities, such as Athens. The phonetic unit set includes segments that are classified as allophones under certain phonological analyses (on the theoretical assumption of additional underlying vowels), because they are phonetically distinctive at the surface level as actually observed. Notably, all palatal consonants and the velar nasal are considered to be distinct segments. The affricates /ʦ/ and /ʤ/ are considered to be single segments and not homorganic doublets (Fourakis et al. 2003; Tzakosta and Vis 2007). The labiodental nasal /ɱ/ is treated as identical to the bilabial nasal /m/ because its pronunciation is optional and nondistinctive.

Phonetic forms have been simplified to remove ambiguity from optional pronunciations and to simplify syllabification. For example, combinations of a nasal consonant followed by a homorganic stop were simplified by dropping the nasal (e.g., all occurrences of /mb/ were converted to /b/). Although, etymologically, it might be considered proper (not obligatory) to pronounce the nasal in certain words and not pronounce it in others, in practice the choice regarding the realization of a nasal is a matter of idiolect and/or social circumstances; phonologically, the nasal is always optional.

Phonetic word forms were automatically syllabified capitalizing on the fact that every full vowel in Modern Greek corresponds to a syllabic nucleus. (Rare exceptions concern a few diphthongs, which cannot be detected automatically, and are effectively ignored in these analyses, treated "normally" as bisyllabic.) Subsequently, the principle of maximal onset (Selkirk 1984) was applied, which states that, cross-linguistically, intervocalic consonants are preferentially assigned to syllabic onsets, i.e., segmental positions preceding syllabic nuclei, rather than syllabic codas, namely, timing slots occupied by segments following syllabic nuclei. In our implementation of this principle, consonant sequences preceding each vowel,

up to the previous vowel or word beginning, were compared against a list of legal onset clusters according to the phonotactic constraints and rules of Greek. This list was created by including all word-initial consonant clusters found in verified word beginnings, manually amended to include a large number of clusters that may be considered legal on phonological grounds (Tzakosta and Karra 2007). Consonant clusters found in the list of phonologically legal clusters were assigned to the onset of the syllable on the right of the boundary. Clusters not found in the list were considered illegal onsets and were therefore broken into a coda and an onset part, spanning two syllables. The maximum right-side subset of the cluster forming a legal onset was assigned to the onset of the following syllable and the left-side remainder was assigned to the coda of the preceding syllable. In this way, onset legality was guaranteed (in accordance with the legal cluster list) whereas coda legality was not.

Finally, orthographic and phonetic forms were aligned at the grapheme-phoneme level, as described in Protopapas and Vlahou (2009), using a list of possible mappings originally based on Petrounias (2002). Orthographic syllabification was derived by joining the graphemes corresponding to the phonetic segments making up each syllable.

3 Indicative results

On the basis of the aforementioned processing, IPLR provides counts of several kinds of units, separately for each corpus. Raw counts include individual letters, phonemes, graphemes, phonetic and orthographic syllables, consonant clusters, and phonetic and orthographic bigrams. Several alternative calculations are offered when there is no generally established or clearly preferable method (e.g., for bigrams). At the word level, IPLR provides counts of phonetic and orthographic neighbors and initial-syllable cohorts. "Standard" neighbors include items of equal length differing in a single position (letter or phone) (Coltheart et al. 1977), i.e., they are single-replacement neighbors. To facilitate research on letter position coding, neighbors differing by a deletion, insertion, or transposition are also provided ("Levenshtein neighbors" of pairwise distance equal to 1; see Yarkoni et al. 2008), as well as stress neighbors, matching the target phonetically from the stressed segment through the end. Word-level calculations include cumulative bigram probability, mean bigram and syllable frequency, and a number of uni- and bi-directional orthographic transparency measures based on Spencer (2009).

Automatic syllabification can only be approximate, due to unresolved phonological issues associated with onset clusters, such as extrametricality, historic relics, recent loans, and morphological considerations. However, the proposed approach, as a rough approximation, facilitates automatic processing and allows calculation of quantities relevant to psycholinguistic research. Our results are consistent with the widely held notion that Greek syllables are predominantly open and relatively simple, as our syllabification resulted in 55.9% of the syllable tokens being CV, followed by V (17.3%), CCV (12.5%), CVC (10.1%), CCVC (2.1%), VC (1.5%), CCCV (.5%) and rare more complex structures (less than 0.1% each).

Detailed results are distributed in freely available spreadsheets at the IPLR site under "Downloads." All counts and calculations are provided for type and token counts separately, as well as taking into account or ignoring stress. Because Greek has lexical stress (Petrounias 2002; Revithiadou 1999) marked in the orthography with a diacritic over the vowel of the stressed syllable, it is possible to consider letters bearing the diacritic as identical to or distinct from the same letters without the diacritic. Similarly, stressed vowel phones can be treated as identical or distinct from the unstressed phones, and this can be important when stress distinctions must be preserved, or when stress itself is the topic of research.

Because the results of all analyses are available online, here we report only a few descriptive statistics, for general reference and information. Table 1 lists type and token statistics from the C corpus, disregarding stress, while Table 2 lists, in order of decreasing token frequency, the letters and phones, and the most frequent phonological syllables, consonant clusters, and letter bigrams. These token counts are arguably more representative of the typical reader's experience with written language than those reported by Ktori et al. (2008) for GreekLex, because the latter were derived using only base forms from a lemma database and therefore exclude many common grammatical forms and the letters and sounds associated with them (see Conrad et al. 2008; Hofmann et al. 2007; Protopapas and Vlahou 2009, pp. 993–994, for further discussion of this issue). Type and token statistics for all measures can be found at http://speech.ilsp.gr/iplr/word_stats.htm, with links to corresponding histograms.

Calculating separately over tokens and over types, each for orthographic and phonetic forms, once taking stress into account and once ignoring it in every case, already presents a daunting selection of individual counts to consider, for what may be the same psycholinguistic variable. The problem is exacerbated for bigrams, where the method of calculating a word-level metric is not standardized; for neighbors, where alternative conceptions have been proposed; and for transparency, where directional and nondirectional alternatives may be based on minimum or mean pairing probability. In IPLR we have taken the option to compute all known alternatives, providing researchers with maximum flexibility. This has the advantage of allowing a close match to designs of studies in other languages, by using the same kind of metric that happens to be available (or chosen) in that language. The disadvantage is that researchers not trying to match a pre-existing design may be left wondering which metric to prefer. We hope that psycholinguistic research will soon produce evidence-based recommendations for specific variables of interest, clarifying the concepts involved in the effects of these variables.

For the moment, we may note that the usual practice of simply averaging bigram counts over a word is mathematically indefensible because it takes each letter into account twice yet fails to account for the base rate of individual letters. Therefore, for bigrams, we recommend preferring the cumulative probability counts, which are calculated using conditional probabilities. For neighbors, new distance metrics based on Levenshtein distance (Yarkoni et al. 2008) seem to be gaining empirical ground over the classic count of Coltheart's *N*. For transparency, Spencer (2010) found that a token-weighted bidirectional index of minimum (rather than mean) transparency accounted for most variance in English-speaking children's early

Table 1 Descriptive statistics for basic units and counts in the C corpus

	Minimum	Maximum	Mean	SD	25%ile	Median	75%ile
Type statistics (N = 206,621)							
Number of letters	1	23	10.07	2.71	8	10	12
Number of phones	1	23	9.45	2.61	8	9	11
Number of syllables	0[a]	11	4.35	1.29	3	4	5
Phonological neighbors[b]	0	51	2.36	3.27	1	2	3
Orthographic neighbors[b]	0	27	1.38	1.58	0	1	2
Token statistics (N = 29,557,090)							
Number of letters	1	23	5.43	3.17	3	5	8
Number of phones	1	23	5.02	3.03	3	4	7
Number of syllables	0[a]	11	2.38	1.45	1	2	3
Phonological neighbors[b]	0	51	10.72	8.42	3	9	17
Orthographic neighbors[b]	0	27	5.88	4.83	2	4	9

[a] The orthographic word forms γι and κι (contracted forms of για and και, respectively) attach themselves phonologically to the following word and do not constitute syllables by themselves.

[b] Coltheart's N

reading accuracy; however, it remains to be investigated which metric is most useful for more developed stages of reading or for accounting for response times and reading fluency, especially in languages with overall more transparent orthographies.

4 Online tools and technical information

IPLR processing is done using functions written in the Python programming language (see http://www.python.org). A special library has been put together as a Python module that can be downloaded and used in custom applications. To ensure maximum transparency and replicability of the methods applied, all code supporting the online calculations and tools is also available for downloading, along with simple examples for use. Additional code is included to perform rule-based graphophonemic transcription, token frequency counting, and a number of auxiliary text-processing functions. All code is provided free of charge and without restriction for noncommercial research or educational use.

In addition to tables of counts and calculations, IPLR provides a set of online tools to help researchers search, identify and evaluate stimuli. These online services are implemented using Active Server Page Extended (asp.NET) applications providing security and flexibility in calling the corresponding text processors.

- The NUM tool provides numerical data for user-provided input. That is, the user types in letter strings (words or nonwords) in the corresponding field and selects the desired quantitative information about them by clicking on the corresponding checkboxes. Details about the nature (and unit) of each measure are provided at http://speech.ilsp.gr/iplr/measures-variables.htm. The user also selects the

Table 2 Sublexical units in order of decreasing token frequency in the C corpus

Letters		Phones		Syllables		Clusters		Bigrams	
α	11.15	i	15.03	a	5.74	st	18.51	ου	2.47
ο	10.27	a	10.41	ɛ	3.92	pɾ	8.57	το	2.26
ι	9.16	o	10.09	i	3.90	ks	5.85	ει	1.92
ɛ	8.93	ɛ	9.24	o	3.45	ft	4.46	τη	1.76
τ	8.31	s	8.91	si	2.91	zm	3.12	αι	1.57
ν	6.30	t	8.37	ti	2.58	tɾ	2.82	πο	1.56
η	5.25	n	6.23	mɛ	2.37	kɾ	2.34	χα	1.55
σ	4.54	ɾ	4.56	na	2.09	ps	2.11	τα	1.53
υ	4.52	p	4.35	po	2.00	pl	2.05	στ	1.29
ρ	4.23	m	3.32	cɛ	1.96	rɣ	2.03	να	1.19
π	4.04	u	2.66	ri	1.84	kt	2.02	με	1.17
κ	3.89	k	2.60	ði	1.80	xɾ	2.01	αν	1.15
ς	3.44	l	2.42	ta	1.69	ðj	1.77	ια	1.11
μ	3.32	c	1.95	to	1.69	ɣf	1.60	ιχ	1.10
λ	2.56	ð	1.86	ka	1.66	ðɾ	1.60	τι	1.09
ω	2.26	f	1.35	li	1.59	vɾ	1.60	ρο	.98
δ	1.72	θ	1.26	ni	1.48	kl	1.48	ων	.90
γ	1.44	v	.82	tɛ	1.43	ɾj	1.47	ης	.84
χ	1.21	ɣ	.78	ci	1.33	sç	1.32	εν	.84
θ	1.17	ç	.76	tu	1.31	ɾç	1.31	ρα	.84
φ	.83	x	.73	pi	1.29	pt	1.23	ερ	.83
β	.59	ɟ	.69	tis	1.29	ɣm	1.12	ρι	.79
ξ	.43	z	.58	ma	1.23	sc	1.06	ατ	.78
ζ	.32	d	.55	no	1.15	vl	1.02	τε	.77
ψ	.14	b	.18	nɛ	1.14	fθ	.98	μα	.76
		ɟ	.07	pa	1.04	sk	.90	οι	.76
		g	.07	ton	1.02	ɾn	.88	ον	.75
		ʦ	.04	mi	1.00	mv	.85	κο	.70
		ɪj	.04	ɾa	.96	sp	.84	σε	.69
		ʎ	.03	pro	.96	sf	.82	ισ	.68
		ɲ	.03	pu	.94	mf	.81	υν	.66
		ʣ	.01	sɛ	.93	nð	.78	απ	.66
				pɛ	.93	ɾm	.75	ες	.63
				ko	.89	xθ	.75	ση	.63
				sti	.82	xn	.74	ην	.62
				tin	.78	stɾ	.72	νο	.61
				lo	.67	ɣn	.68	πρ	.59
				la	.67	ɾt	.68	δι	.59
				θi	.60	fs	.59	πα	.58
				ɾo	.59	pç	.57	αρ	.57

Frequencies are per cent of the total corresponding units in the C corpus, calculated over word tokens, ignoring stress and diacritics (accent and diaeresis). For phonetic syllables, consonant clusters, and letter bigrams, only the 40 most frequent are shown

corpus on which the results should be based, whether stress should be taken into account or ignored, and whether the result should be displayed on the web browser or returned as a downloadable file.

- The TXT tool complements the quantitative information provided by the NUM tool with text results such as syllabification, alignment, and neighbor sets. Here, the user types in a single letter string and selects the types of information desired.
- The SEL tool selects words based on user-provided numerical criteria on the aforementioned counts, allowing researchers to retrieve words with specific properties. The user may fill out minimum and/or maximum values for as many of the available measures as desired, and selects the corpus to be searched and whether to ignore stress. To help users set criteria and evaluate the results returned, the distributions of all available counts are provided in tabular and graphical form at http://speech.ilsp.gr/iplr/word_stats.htm.
- Finally, the FIND tool identifies words matching a user-provided pattern, including wildcards for arbitrary characters. The matching pattern may be either orthographic—a letter string—or phonetic—a phone string,—allowing retrieval of words with specific letter or phone sequences. The phonetic notation used in IPLR is based on the Latin alphabet, in order to avoid complications arising from encoding and font selection incompatibilities, allowing easy keyboard input, and can be found at http://speech.ilsp.gr/iplr/PhoneticSymbols.htm.

5 Limitations

Limitations to the application and exploitation of IPLR data arise due to the source of the materials and the phonetic transcription. Because all IPLR calculations are based on written text corpora the results may not be representative of the spoken language and may be generalized with great caution. Moreover, despite the substantial diversity in origin of the text included in C corpus (based on the HNC), there is always a concern about representativeness for specific genres, populations, or applications, even when investigating written language. Furthermore, restriction of analyses to individual isolated word types necessarily misses information related to phrasal context and interword relations and interactions, which may be relevant for some types of investigations.

With respect to the phonetic calculations, it should be kept in mind that all information is derived from the transcriptions of an automatic text-to-speech system and are not based on recordings of native speakers. This particular transcription system is known to produce highly accurate and intelligible speech, however this does not mean that the output is an accurate approximation of what might be computed on the basis of acoustic-phonetic analyses of actual speech material. Even taking into account the unavoidable dialectal restriction to the "standard Greek" spoken in major cities, pronunciation nuances and variation among and within words are lost, as are possible phonological effects across words. Therefore, it is advisable to consider additional pretesting and validation of selected materials, when working with oral language, instead of exclusive reliance on the IPLR metrics.

6 Conclusion and future directions

In sum, IPLR is a resource addressing a significant need for researchers interested in the Modern Greek language, which can now be offered thanks to the availability of text corpora and a small number of key text and speech technologies available at ILSP. Further enhancements will be made as needed, incorporating more measures and tools whenever possible. Perhaps we are approaching the day of international collaboration for the standardization of processing and measurement and for the unification of multi-language corpora to facilitate cross-linguistic work based on comparable information. Integration with phonetic analysis and additional tools from the natural language processing field is also highly desirable. It is anticipated that this resource will greatly enable future psycholinguistic work in Greece and abroad.

Acknowledgments We thank Despina Paizi for information on Italian databases and Eleni Orfanidou for comments on the manuscript.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K., Kessler, B.,Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445–459

Barca, L., Burani, C., & Arduino, L. S. (2002). Word naming times and psycholinguistic norms for Italian nouns. *Behavior Research Methods, Instruments & Computers, 34*, 424–434.

Chalamandaris, A., Raptis, S., & Tsiakoulis, P. (2005). Rule-based grapheme-to-phoneme method for the Greek. In *Inter speech 2005* (pp. 2937–2940).

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Conrad, M., Carreiras, M., & Jacobs, A. M. (2008). Contrasting effects of token and type syllable frequency in lexical decision. *Language and Cognitive Processes, 23*, 296–326. doi:10.1080/01690960701571570.

Fourakis, M., Botinis, A., & Nirgianaki, E. (2003). Hroniká haraktiristiká ton simfonikón akolouthión [ps], [ts], kai [ks] stin ellinikí [Temporal characteristics of [ps], [ts], and [ks] consonant sequences in Modern Greek]. In *Proceedings of the 6th international conference on Greek linguistics (ICGL6), Rethimno, Greece, September 18–21*.

Gomes, I., & Castro, S. L. (2003). Porlex, a lexical database in European Portuguese. *Psychologica, 32*, 91–108.

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., et al. (2000). Design and implementation of the online ILSP corpus. In *Proceedings of the 2nd international conference of language resources and evaluation (LREC)* (Vol. 3, pp. 1737-1740), Athens, Greece.

Hofmann, M. J., Stenneken, P., Conrad, M., & Jacobs, A. M. (2007). Sublexical frequency measures for orthographic and phonological units in German. *Behavior Research Methods, 39*, 620–629.

Ktori, M., van Heuven, W. J. B., & Pitchford, N. J. (2008). GreekLex: A lexical database of Modern Greek. *Behavior Research Methods, 30*(3), 773–783.

Lambert, E., & Chesnet, D. (2001). Novlex: Une base de données lexicales pour les élèves de primaire. *L'Année Psychologique, 101*, 277–288.

Laudanna, A., Thornton, A. M., Brown, G., Burani, C., & Marconi, L. (1995). Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente. In S. Bolasco, L. Lebart, & A. Salem (Eds.), *III giornate internazionali di analisi statistica dei dati testuali* (Vol. I, pp. 103–109). Roma, Italy: Cisu.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments, & Computers, 36*, 156–166.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers, 36*, 516–524.

Peereman, R., Lété, B. B., & Sprenger-Charolles, L. (2007). Manulex-Infra: Distributional characteristics of grapheme-phoneme mappings, infra-lexical and lexical units in child-directed written material. *Behavior Research Methods, 39*, 579–589.

Petrounias, E. V. (2002). *Neoellinikí grammatikí kai sigkritikí análisi, tómos A: Fonitikí kai eisagogí sti fonología* [Modern Greek grammar and comparative analysis, Vol A: Phonetics and introduction to phonology]. Thessaloniki: Ziti.

Protopapas, A. (2006). On the use and usefulness of stress diacritics in reading Greek. *Reading & Writing: An Interdisciplinary Journal, 19*(2), 171–198.

Protopapas, A., & Vlahou, E. L. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods, 41*, 991–1008

Revithiadou, A. (1999). *Headmost accent wins: Head dominance and ideal prosodic form in lexical accent systems*. LOT Dissertation Series 15 (HIL/Leiden University). The Hague: Holland Academic Graphics. (Available from http://www.roa.rutgers.edu)

Rinaldi, P., Barca, L., & Burani, C. (2004). A database for semantic, grammatical and frequency properties of the first words acquired by Italian children. *Behavior Research Methods, Instruments & Computers, 36*, 525–530.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Spencer, K. A. (2009). Feedforward, -backward and neutral transparency measures for British English. *Behavior Research Methods, 41*, 220–227.

Spencer, K. A. (2010). Predicting children's word-reading accuracy for common English words: The effect of word transparency and complexity. *British Journal of Psychology, 101*(3), 519–543. doi:10.1348/000712609X470752.

Tzakosta, M., & Karra, A. (2007). A typological and comparative account of CL and CC clusters in Greek dialects. In *3rd International conference on Modern Greek dialects and linguistic theory, University of Cyprus, June 14–16*.

Tzakosta, M., & Vis, J. (2007). Phonological representations of consonant sequences: the case of affricates vs. 'true' clusters. In *8th International conference on Greek Linguistics, University of Ioannina, August 30–September 2*.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979.