Dialectologia 8 *(2012), 1-26.*
ISSN: 2013-2247

Received 18 May 2011.
Accepted 10 July 2011.

MODELING PHONETIC VARIATION IN PLURICENTRIC LANGUAGES: AN INTEGRATIVE APPROACH[1]

Simone ASHBY[*], Mário Eduardo VIARO[±], Sílvia BARBOSA[*], and Neuza CAMPANIÇO[*]
Instituto de Linguística Teórica e Computational (ILTEC), Lisbon, Portugal[*]
Universidade de São Paulo, São Paulo, Brazil[±]
[*]{simone, silvia, neuza}@iltec.pt; [±]maeviaro@usp.br

Abstract

Adaptive speech technologies offer a vehicle for representing pluricentric language variation and the description of both dominant and non-dominant speech varieties. In this article, the work of the LUPo project is described for modeling phonetic variation across national and sub-national varieties of the Portuguese language. While the motivation for this research is based around the development of high-quality pronunciation lexica for a Portuguese text-to-speech system – a goal which, itself, is aimed at facilitating the entry of lesser or undocumented variants into the digital domain – the repercussions for pluricentricity are far reaching. We describe how systems such as LUPo can be used to model variation across phonetically similar and disparate national, sub-national, and sociolectal varieties, as well as presenting linguists with a means of testing and observing notions of linguistic distance in terms of shared or innovative rules and phonetic features, and for evaluating the pulling effect of different linguistic centers.

Key words

pluricentric language variation, Portuguese, phonetics, pronunciation generator, dialectometry

[1] This article is an expansion of a paper that was printed in the *Proceedings of the International Conference on Pluricentric Languages* (2010).

LA REPRESENTACIÓN DE LA VARIACIÓN FONÉTICA EN LENGUAS PLURICÉNTRICAS. UNA APROXIMACIÓN INTEGRAL

Resumen

Las tecnologías de adaptación del habla ofrecen un vehículo para representar la variación lingüística pluricéntrica y la descripción de variedades de habla dominantes y no dominantes. En este artículo, se describe el trabajo que lleva a cabo el proyecto LUPo para la representación de la variación fonética en las variedades nacionales y subnacionales del portugués. Mientras que la motivación para esta investigación se basa en el desarrollo de léxicos de pronunciación de alta calidad para un sistema de texto a voz del portugués − un objetivo que, en sí mismo, tiene por objeto facilitar la entrada en el dominio digital de las variantes menores o indocumentadas − las repercusiones para pluricentricidad son de largo alcance. Describimos cómo una sistema como LUPo pueden representar la variación a través variedades nacionales, subnacionales y sociolectales fonéticamente similares y diferentes, de la misma manera que los lingüistas pueden utilizarlo con un medio de prueba y observación de las nociones de distancia lingüística en términos de reglas compartidas o innovadoras y de rasgos fonéticos, y para evaluar el efecto de extracción de los diferentes centros de lingüística.

Palabras clave

variación lingüística pluricéntrica, portugués, fonética, generador de pronunciación, dialectometría

1. Introduction

This work is a description of the LUPo online interface, as well as a presentation of results for observing pronunciation varieties from Brazil, Mozambique, and Portugal. This work marks the first phase of a three-year research project dedicated to the creation of an accent-independent lexicon and rule system for generating accent-specific pronunciations for regional variants of Portuguese. More in-depth information about the original Unisyn Lexicon upon which LUPo is based can be found in Fitt (2000).

The motivation for this research is based around the development of high-quality pronunciation lexica for a pan Lusophone text-to-speech system. As speech technologies become an increasing part of our everyday lives, the users of these technologies represent an ever widening speaker base. Adapting such technologies to a wider number of speakers — and *topolects* — and representing countries and regions for whom such development concerns have been largely overlooked carries significant

economic and political weight in narrowing the global digital divide, and promoting further research among lesser studied varieties.

Through the establishment of a linguistically derived rule system for the explicit treatment of allophones within and across regional varieties, LUPo circumvents the cost of producing high-quality phonetic transcriptions by hand, while attracting a wider pan Lusopohone audience to the online lexical database in which it resides, and providing the research community with a vast resource of Portuguese accent data for evaluating speech applications and testing diachronic, phonological, and sociolinguistic theories.

Here, a seminal effort is presented towards developing systematized, multiple, complete phonetic models for non-standard varieties of Portuguese as it is actually spoken in different parts of the world. Broad phonetic segmental models [2] were developed based on idiolectal data representing Belém, of the northeastern coast in Pará, Brazil, and the capital city of Maputo, in Mozambique. Similarly, broad phonetic models of the standard Lisbon and São Paulo accents were developed based on descriptions of these varieties in the literature, along with the help of dictionary pronunciations and native speaker insights.

This work provides a window into the segmental models for the above idiolects, as contrasted with those for the standard Lisbon and São Paulo accents. A selection of post-lexical rules is presented, along with a description of how one of LUPo's key components, the regional accent hierarchy, enables the sharing of rules across pluridimensional dialectal and sociolectal varieties. Finally, a description is provided of the LUPo system as it currently exists, and some preliminary results are presented for observing and comparing national and sub-national varieties of the Portuguese language through an analysis of shared rules and the application of the Levenshtein distance algorithm (Levenshtein, 1965]).

2. Background

Portuguese is a pluricentric language spoken by one-fifth of the world's population, and with regional variants spanning Africa, Asia, Europe, and South

[2] The LUPo project also aims to treat cross-word phenomena, such as external sandhi. Acoustic modeling and suprasegmental feature descriptions will be undertaken in the synthesis project to follow.

America.[3] Portuguese is a recognized official language in Angola, Brazil, Cape Verde, East Timor, Equatorial Guinea, Guinea-Bissau, Macau, Mozambique, Portugal, and São Tomé and Principe.

Significant lexical, grammatical, phonological, and phonetic differences distinguish what are recognized as the language's primary centers: Brazil and Portugal. It is assumed, at least in Portugal, that Luso-African and Luso-Asian varieties follow the standard European accent, i.e. the variety of Portuguese spoken in Coimbra and Lisbon. However, as these non-dominating varieties become more established and widespread in countries such as Mozambique, home-grown Portuguese speech varieties are emerging with their own lexicons, grammars, morphologies, and phonologies (Gonçalves 2010: 14; Lopes 1999: 122; Baxter 1992: 23-27). One of the principle aims of this article is to open the way towards examining both dominating and non-dominating regional varieties of the Portuguese language, and establish an initial inquiry into the manner and extent to which non-standard varieties from Africa and Brazil differ from respective dominating varieties, i.e. the European Portuguese (EP) and Brazilian Portuguese (BP) standards.

2.1. Data selection[4]

The idiolects and standard varieties selected for this study represent cities from three of the four continents where Portuguese has an official status, i.e. Africa, Europe, and South America. It should be noted that the selection of data presented is a reflection of the work performed during LUPo's first year, and that data collection and analysis are *ongoing* as part of an overall effort to describe and model 10 or more regional variants across a wide global distribution. Our main objective in the setup of the current study was the inclusion of a Luso-African spoken variety, for which there is extremely little published research, plus the inclusion of two sub-national varieties — Belém and standard São Paulo — for the purpose of demonstrating LUPo's regional accent hierarchy (see section 3.2.2). The focus on idiolects was a pragmatic decision, based on a preference for comparing complete segmental models. Ultimately, topolectal varieties

[3] See Baxter (1992) for a general discussion of Portuguese as a pluricentric language.
[4] Details concerning the data collection and modeling of idiolects are provided in section 3.1.

presented within the LUPo system will be derived from composite models, combining data observations from multiple informants, along with findings from relevant variationist studies.

The informants selected for this study were recorded in Lisbon, Portugal. Both are male, ranging in age from 30 to 36, and consider Portuguese to be their mother tongue. At the time of recording, the informant from Belém (IB01) had been residing in Lisbon for more than four years, while the Mozambican informant (IM01) had only just arrived in Portugal. Given these circumstances, and the fact that LUPo's data elicitations are conducted by researchers from Lisbon, dialectal *accommodation*, or "adjustments in pronunciation and other aspects of linguistic behavior in terms of a drive to approximate one's language to that of one's interlocutor" (Trudgill 1983: 143) should be considered a potential factor affecting IB01's dialect of origin. However, a careful analysis of IB01's phonetic characteristics, as partially exemplified in sections 3.3 and 4) and responses by this informant to LUPo's attitudinal questionnaire do not appear to lend evidence to a convergent contact situation.

2.2. Dialectological background

2.2.1. Belém and coastal Pará, Brazil

After neighboring Amazonas, the northern Brazilian state of Pará is the second largest state in Brazil in terms of land mass. Belém, which is situated along the banks of the Amazon estuary, is the second most populous city in Brazil's northern region. The most recent estimate from Brazil's Institute of Geography and Statistics indicates a population of 2,335,000 people residing in the greater Belém area (IBGE, 2010).

As with the other Brazilian states, Portuguese is the official language of Pará and that which is primarily taught in schools. State sponsored bilingual education programs exist for a handful of surviving indigenous languages, but these are largely relegated to the outlying rural areas where indigenous communities are concentrated. A number of other European and Asian languages, such as German, Italian, and Japanese, are maintained by Pará's immigrant population.

In general, the Portuguese language varieties evident in radio and television broadcasts from Brazil's two largest urban centers, Rio de Janeiro and São Paulo, are regarded as the country's prestige dialects. However, there is little evidence in the existing literature to define the specific attitudes and preferences held by Portuguese language speakers living in Pará concerning prestige varieties.

The regional Portuguese dialect, *Paraense*, has been the subject of a fairly large number of variationist studies, many of which were initiated as part of the *Atlas Linguístico do Brasil* (ALiB) project and its phonetic counterpart for the state of Pará, the *Atlas Linguístico Sonoro do Pará* (ALiSPA) project. Those studies dedicated to describing the accent of Belém, either specifically or in part, include work by: de Carvalho (2000) and Scherre & Macedo (1991) concerning variable realizations of post-vocalic /s/; Lopez (1979), citing variable realizations of /r/, /l/ and /s/ in syllable-final position, while in word-final position, these consonants were found to adhere to external sandhi rules; Vieira (1983) concerning variable realizations of the alveo-dental fricatives /s/ and /z/ both word finally and preceding a voiceless consonant; Oliveira & Razky (2010), who report a high rate in the realization of pre-vocalic /l/ as the palatal lateral [ʎ] before the high vowel [i]; Brandão & Cruz (2005), confirming the existence of the open vowels /ɛ/ and /ɔ/ in unstressed positions; and Nina (1991), whose study of pre-tonic /e/ and /o/ shows an assimilative tendency by speakers to produce raised or lowered tokens in accordance with the height of the vowel in the following syllable.

2.2.2. Mozambique and its capital, Maputo

Mozambique extends along the Indian Ocean, from its northern border with Tanzania to the country's southwest reaches, bordering Swaziland and South Africa. The interior is made up of horizontally striated river valley settlements that extend from the much larger urban areas that dot the coast. At the time of writing, the population of Mozambique was estimated at over 22 million, with 37% of the population residing in cities (CIA 2010). The capital city of Maputo is located in the country's southernmost tip, an area that is integrally connected with South Africa in terms of a shared economic structure and communications network (Newitt 2002: 186).

Mozambique, like many other African countries, is home to a linguistically diverse population. The vast majority of Mozambicans speak one of a variety of indigenous Bantu languages as their mother tongue. During the 1960s, in its war for independence from Portugal, leaders of the resistance adopted Portuguese as a means of uniting nationalist freedom fighters across the country's diverse language topography. To this day, Portuguese remains the official language of Mozambique, where it is spoken as a lingua franca by 33% of the population, an additional 6.5% of which regard Portuguese as their native language (Gonçalves 2010: 26). Portuguese is used in all official communications. It is the language of instruction in Mozambican schools and the Eduardo Mondlane University, and it is used by the majority of Mozambican media outlets.

Soon after 1975, when Mozambique achieved independence from Portugal, lawmakers and educators determined that the teaching of Portuguese in schools should aim towards EP. However, in subsequent years, "practice showed that such an idealistic goal was not achievable, and even no longer desired because it lacked the marks of an emerging national identity" (Lopes 1999: 123). Since then, Mozambique has exercised what Lopes (1999: 123) describes as a "*laissez-faire* policy" concerning the normativization and standardization of Portuguese. Meanwhile, the status of Portuguese in Mozambique has increasingly come to be regarded as a language under threat due to the strengthening of economic ties with South Africa and Mozambique's other anglophone neighbors, its recent entry into the British Commonwealth, and economic and linguistic intervention from France (da Conceição 1999: 22).

Nevertheless, Portuguese retains its official status in Mozambique and represents an indispensable tool for communicating outside the family and enhancing upward social mobility. Portuguese has also been increasingly appropriated as a means of expression by writers and musicians.

> The alterations to the Portuguese language reveal a logic that goes well beyond the linguistic domain, and translate a different world view and lifestyle. The Mozambicans are in the process of transcending their role as simply users of the Portuguese language and assuming a status in which they are co-producers of this means of expression[23] (Couto 1986).

To date, much of the work of describing Mozambican Portuguese has been focused on descriptions of its lexical and syntactic features (e.g. Carvalho 1991; Chimbutane 1998; Dias 2009; Diniz 1988; Gonçalves 1986; Issak 1998; Lopes 1979; Machungo 2000; Maciel & Pascoal 2002). Extremely little attention has been devoted to describing the different phonetic features evident in varieties of Mozambican Portuguese. Gonçalves (1986) offers an account of what were previously regarded as "deviations" from EP produced by Portuguese speakers in the Maputo area, while Gonçalves (2010) focuses on the country's multilingual composition and language contact effects on local varieties of Portuguese. In the latter study, Gonçalves presents just a few examples illustrating trace effects of local Bantu phonologies on spoken varieties of Mozambican Portuguese, citing: a tendency among native speakers of Macua for the voiced obstruents /b/, /d/, and /g/ to be realized as the voiceless counterparts [p], [t], and [k]; use of the uvular trill [R] among native Changana speakers (originally reported in Sitoe & Ngunga 2000); and an overall preference for open syllables.

2.2.3. The Lisbon and São Paulo standard varieties

The most comprehensive study to date detailing the standard Lisbon accent is by Mateus & d'Andrade (2000). While the focus of this treatment concerns generative accounts of the EP phonological system, phonetic segmental realizations are presented with considerable attention paid to their current usage. As such, this proved an indispensable resource in developing our segmental model of the standard Lisbon accent. Cagliari's (1981) dissertation on BP phonetic features contains one of the more detailed descriptions of the standard São Paulo variety (also known as paulistano). This, and the input of this study's native paulistano co-author form the basis of LUPo's standard São Paulo segmental model.

3. LUPo

As indicated in section 1, LUPo provides the basis for a subsequent project aimed at developing a text-to-speech (TTS) system that is capable of generating synthetic speech from text for multiple regional variants of Portuguese. This is an important direction for speech technology, given that most TTS systems are built using data from a single accent, usually what is considered to be the standard variety for a given language. Instead of expending thousands of man hours to transcribe a complete dictionary for just one accent, our methodology involves a careful modeling of the accent's sound system. This information is interpreted as a set of rules, which are applied to a accent-independent lexicon (i.e. a list of words with their metaphonemic representations) for generating accent-specific phonetic transcriptions. In this way, LUPo succeeds in dramatically reducing the investment spent per regional variety, while yielding high-quality pronunciation output.

In addition to serving as the input to a TTS system, we are also developing a Web interface for the general public via the existing Portal da Língua Portuguesa online lexical knowledge base (http://www.portaldalinguaportuguesa.org). Users will have the option of selecting from a range of dialects in which to display the pronunciation for a given word. Inclusion of LUPo in the Portal will enhance the Portal's value as a pan Lusophone resource and the only one of its kind to provide detailed and varied phonetic output for a large number of Portuguese dialects. Indeed, it will be the first freely available online resource to provide any manner of high-quality transcription data for Portuguese.

3.1. Data collection, modeling, and evaluation

In general, the collection and modeling of accent data involves using multiple means — from published studies and corpora (labeled or otherwise), to the use of linguistically trained informants, to the collection and analysis of new speech data, to the use of pronunciation lexica (in the case of standard varieties) — to construct complete segmental models for spoken variants of Portuguese. For each accent or idiolect treated, a complete segmental model consists of: a long list of

morphophonological contexts (especially those which are most vulnerable to change) and their corresponding phonetic realizations, i.e. a set of morphophonological post-lexical rules; conditions for the ordering of rules; and a list of lexical exceptions.

Materials for the elicitation of read speech are based on those established in Rodrigues (2003), with the inclusion of a small set of additional words and phrases deemed necessary for capturing other relevant contexts. Audio recordings and stimulus prompts are controlled by a researcher, who remains seated in the same room as the informant, albeit in the periphery and not directly in front of the informant. For the read speech elicitation task, informants are asked to read the individual phrases and sentences projected in front of them on PowerPoint slides. When this task is completed, the elicitation of spontaneous speech data is conducted in the form of an oral questionnaire for obtaining general speaker information and attitudinal data. Recordings are done using a Marantz digital voice recorder, with a microphone positioned on the table in front of the informant. Later, the roughly 90-minute-long digital audio files are separated into recording blocks.

Corpus-based accent models are then developed through the assessment of segmental data by trained phoneticians, who use Praat (Boersma & Weenink 2010) to identify and label target segments, based on a combination of auditory judgment and waveform and spectrogram analysis. Each accent model undergoes a separate pass by a total of three phoneticians until agreement is reached concerning the complete set of data points described. Accent models, corresponding post-lexical rules, and LUPo output transcriptions are further subjected to an external review by a linguistically trained native speaker.

3.2. System architecture

LUPo's core components include: an accent-independent master lexicon of underspecified pronunciations (including part of speech and frequency information), a regional accent hierarchy, an exceptions dictionary, and the application (through Perl scripts) of morphophonological rules that transform the master lexicon pronunciation into the target output (Figure 1).

Dialectologia 8 *(2012), 1-26.*
ISSN: 2013-2247

3.2.1. Master lexicon

LUPo's accent-independent lexicon, or *master lexicon*, consists of entries formed from an extended set of X-SAMPA-based key symbols that capture a rough approximation of each entry's underlying phonological form. These can best be understood as 'metaphonemes', and take from the ideas in Wells (1982) as a means of "[a]bstracting away from phonetics [so] that a single lexicon can represent numerous different accents" (Fitt & Isard 1999: 823). Key symbols also allow for the inclusion of morphology, along with stress and syllable boundary information. For example, encoding morphological boundaries in the non-hyphenated word compounds *coigual* and *coutente* enables LUPo to properly interpret the contiguous vowel sequences /oi/ and /ou/ as contexts for hiatus, instead of erroneously interpreting these sequences as diphthongs. To illustrate, master lexicon entries for the above words appear roughly as follows, with double angle brackets demarcating the bound prefix morpheme <co->:

(1) coigual adjectivo mf < k_c o < . { i . g w_u "a 5_l }

(2) coutente adjectivo mf < k_c o < . { u . t "e~ . t i_e }

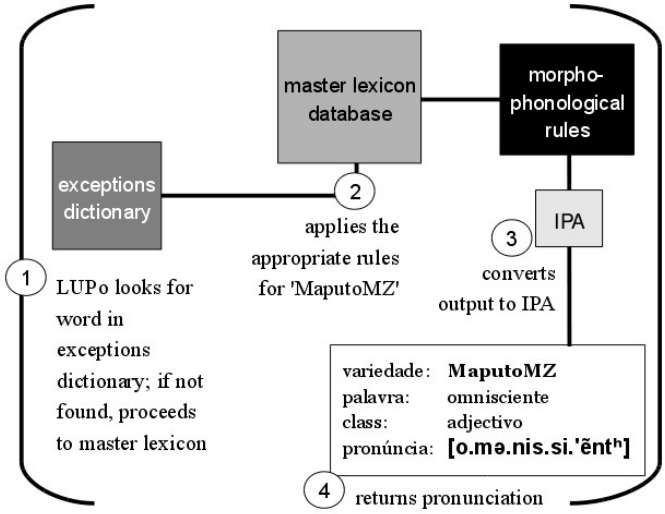

Figure 1. LUPo's architectural components

Note that in examples (1) and (2) above, curly brackets are used to describe free morphemes, while an underscore symbol is used for linking metaphonemes (shown in X-SAMPA symbols) with their non-identical graphemic counterparts.

While the pilot version of the master lexicon used in this study is restricted to 125 word forms, the full version will eventually contain metaphonemic entries for every lemma in the Portal da Língua Portuguesa (hereafter referred to as the *Portal*). For the current pilot study, we included inflected forms, such as *beberam*, *consumidores*, *unicamente*, and *pedrazinha* as a means of evaluating pronunciation effects conditioned by morphology. In subsequent versions, only lemmas will be stored in the master lexicon, which will draw from the lexically rich Portal infrastructure to capitalize on the inflectional and derivational links, spelling variants, part of speech information, foreign loan word and toponym attributes, and morphological information currently contained therein. In this way, LUPo will be capable of generating transcriptions for the words *actividade* and *practicamente* without the need to store these and other inflected forms in the master lexicon.

3.2.2. Regional accent hierarchy

As with LUPo's other components, the model for the regional accent hierarchy is based on that of the original English Unisyn Lexicon (Fitt 2000), and is made up of a system of files containing variant specifications and rule scores. An example from Mozambique is presented in Figure 2. The first set of lines is an entry in the file 'lupo_towns', with 'map' representing the capital city of Maputo, and the next set of abbreviations representing a system of levels that correspond to COUNTRY, REGION, TOWN, and PERSON. The subsequent set of lines is taken from a file called 'lupo_scores', wherein a general rule is attributed at the town 'TWN' level for the simplification of the nasal diphthong /e~j~/ as the monothong nasal vowel [e~] across varieties from both Maputo and Belém. Note the different rule scores ('1' and '2') assigned to each topolect, which, in the case of Belém, restricts the application of this rule to just non-tonic contexts.

What is interesting about this hierarchical system is the inheritance by each node of features from the previous node, provided the inheritance is not broken by the

introduction of a competing feature (or features) at a lower level. As the lowest level in the hierarchy, rules attributed at the person 'PER' level override competing specifications from all the higher levels. By adding features at the PERSON level, we can characterize a mesolectal variety of young urban speakers, or even that of an individual — say Mozambique's current president Armando Emílio Guebuza — while implicitly treating the remaining set of allophones as inherited from the upper nodes TOWN, REGION, and COUNTRY.

3.2.3. Exceptions dictionary

Economy underscores virtually all of the components of the LUPo model, including its list of exceptions, which need only be expressed for root forms, given a means of generating derived and inflected forms. When local exceptions are found, they are added to LUPo's exceptions dictionary, with links to the regional accent hierarchy for specifying to which variant they belong.

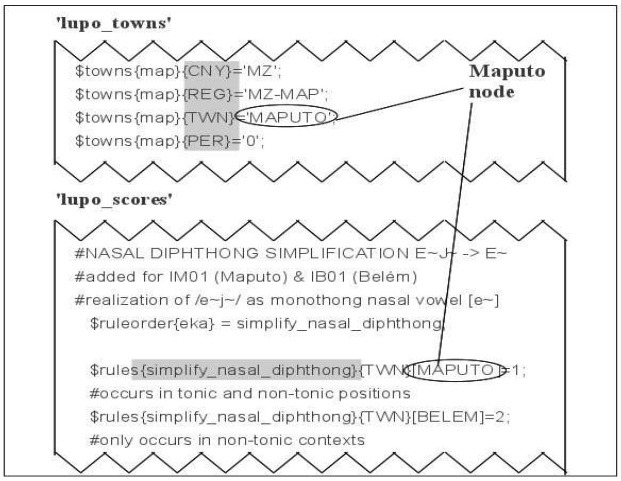

Figure 2. Extracts from LUPo's regional accent hierarchy

3.2.4. Rule system

LUPo stores allophonic rule sets that exploit morphological boundaries to express different accent-specific rules, most of which are post-lexical. Similar to the justification presented for the design of LUPo's master lexicon (section 3.1), the

representation of morphology in LUPo's pronunciation rules enables the system to identify the correct pronunciation in opaque orthographic contexts, such as the assignment of EP vowel height in the lexically related word pairs *m*[o]*lho* 'sauce' and *m*[u]*lhada* 'wet', and *m*[ɔ]*lho* 'bundle of twigs, sticks, or branches' and *m*[ɔ]*lhada* 'an assembly of people or things.' Given LUPo's direct access to the derivational relationships that are already explicit in the Portal, the post-lexical rules responsible for generating *m*[u]*lhada* from its lexical root *m*[o]*lho* and *m*[ɔ]*lhada* from the corresponding root *m*[ɔ]*lho* realize their effectiveness through a restricted application to morphologically related words.

Perl scripts are used to reference the geographic relationships and rule scores contained in the regional accent hierarchy, and to apply rules to the metaphonemic forms contained in the master lexicon for the generation of accent-specific output. A closer look at the rules is presented in sections 3.3 and 4.

3.3. How it works

General users will soon be able to access LUPo via the Portal da Língua Portuguesa website to select from a list of available topolects and generate accent-specific pronunciations. While this capability is currently restricted to lemmas and a very small number of inflected words, LUPo will ultimately be extended to handle word forms and multi-word texts. With LUPo's online interface, users can select from one of the four accents that have been modeled so far and query the system for the pronunciation of a given word, as demonstrated in Figure 3.

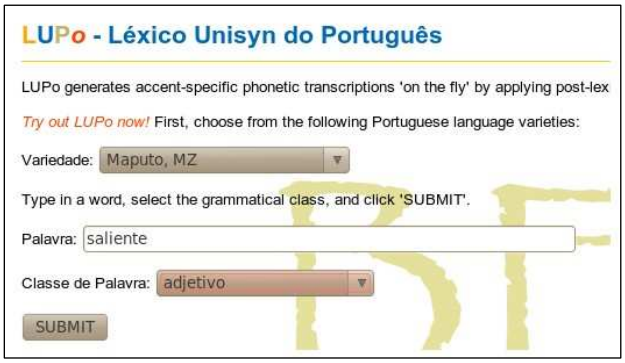

Figure 3. LUPo online prototype

In Figure 4, the result is displayed for IM01 for the adjective *saliente*. Here, one may observe that the syllable onset /l/ is realized as the velarized coda [ɬ], joining the rhyme of the previous syllable, when followed by a rising diphthong, in this case [jẽ]. Note that this speaker inserts the homorganic nasal [n] between the preceding nasal vowel [ẽ] and the following alveo-dental obstruent /t/. Further, the word-final vowel, realized in other contexts by this speaker as [i] and sometimes [ɨ], combines with the preceding obstruent to form the aspirate [ʰ].

Figure 4. Pronunciation of *saliente* by IM01 (Maputo)

A quick comparison with the result for IB01 (Figure 5) shows that this speaker produces a lateral approximant and glide cluster, while the well known BP phenomenon for realizing alveolar obstruents followed by the high vowel [i] as affricates can be observed in the final syllable.

Figure 5. Pronunciation of *saliente* by IB01 (Belém)

The output for the standard São Paulo variety resembles the previous example for IB01 in all but one respect, whereby for the former variety, gliding is realized after the nasal vowel [ẽ].

Figure 6. Pronunciation of *saliente* in the standard São Paulo variety

Output for the standard Lisbon accent (Figure 7) reveals the sort of vowel reduction characteristic among this variety's unstressed syllables, with /a/ in the first syllable reduced to [ɐ], and reduction of the final vowel to [ɨ].

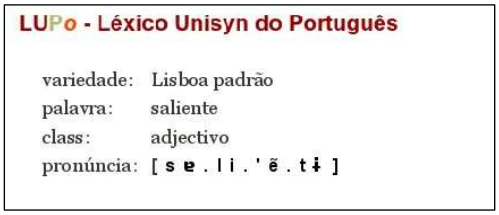

Figure 7. Pronunciation of *saliente* in the standard Lisbon variety

The specific rules applied in the generation of LUPo's accent-specific output are printed in the lower half of the results page (Figure 8). These are not phonological rules in the strict sense, but rather the transformations the master lexicon entry had to undergo to become the sort of output displayed in Figures 4, 5, 6, and 7 above. At the bottom of the page, descriptions of all the rules for a given variety are presented in plain language to make them easier for general users to understand.

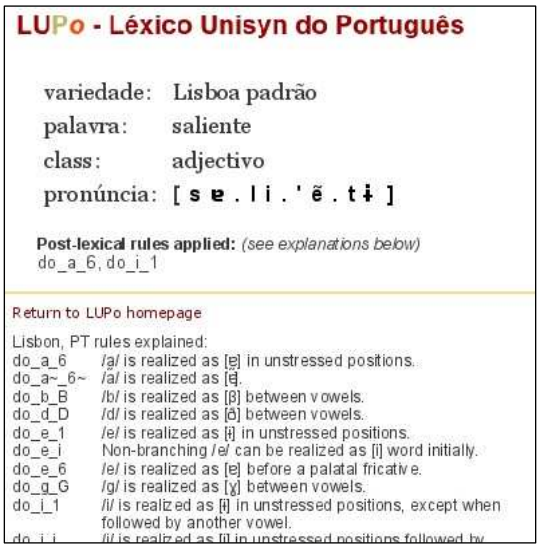

Figure 8. Rule descriptions

The current version of LUPo is designed to display a single variant form for the accent selected. Subsequent versions will display alternate possible forms where

relevant, with a means of identifying which form is more common. Thus, for the standard São Paulo variety, LUPo will be capable of generating two output transcriptions for *rolo*, to accommodate for the realization of word-initial /r/ as the voiceless velar fricative [x] or the voiceless glottal fricative [h].

4. Variant comparisons

In the future, LUPo's online interface will provide users with a comparison module for observing results from more than one variant at a time. For the current study, a cursory set of comparisons were achieved by evaluating shared rules, and performing dialectometric comparisons on output strings through the use of the Levenshtein distance algorithm. Neither should be regarded as a comprehensive analysis of the phonetic similarities and differences describing the four varieties examined in this study. Rather, we present these very preliminary results for the purpose of demonstrating the potential utility of a system such as LUPo for evaluating multiple phonetic data sets and observing the pulling effect of different centers.

4.1. Rule set comparisons

Figure 8 above shows a partial list of rules applied for the standard Lisbon variety. This is a slightly modified list, based on the output of LUPo's conversion script, which prints accent-specific transcriptions along with the relevant set of rules applied. A behind-the-scenes comparison of the different post-lexical rules applied across varieties can thus be easily achieved by converting the aggregate data into a table.

Table 1 shows a subset of rules for effecting reduction, simplification and elision conversions. Note that the rules presented in the first column use X-SAMPA symbols to describe sounds. The use of Roman numerals at the end of a rule indicates that its applicability is tied to multiple sets of conditions. For example, the rule 'do_r_4_II' converts the metaphonemic symbol /r/ to the alveolar flap [ɾ] syllable initially between vowels, while 'do_r_4_III' performs the same conversion within consonant clusters.

rule	std Lisbon	std São Paulo	IB01 (Belém)	IM01 (Maputo)
delete_final_i_resyllabify				**X**
delete_final_r		**X**	**X**	
delete_initial_e_resyllabify			**X**	
denasalize			**X**	
do_5_w		**X**	**X**	
do_a_6	**X**			**X**
do_e_1	**X**			
do_e_6	**X**			
do_e_h_resyllabify				**X**
do_ej_6j	**X**			
do_i_1	**X**			**X**
do_initial_1	**X**			
do_initial_i	**X**			**X**
do_initial_I			**X**	
do_nasal_6	**X**			
do_nasal_i		**X**		
do_nasal_schwa		**X**	**X**	
do_o_u	**X**			
do_r_4_I	**X**	**X**		**X**
do_r_4_II	**X**	**X**	**X**	**X**
do_r_4_III	**X**	**X**	**X**	**X**
do_r_4_IV	**X**	**X**	**X**	**X**
do_schwa_I		**X**	**X**	
do_schwa_II			**X**	
do_tonic_6	**X**			
do_u_U		**X**	**X**	**X**
simplify_affric_seq			**X**	
simplify_nasal_diphthong			**X**	**X**
simplify_vowel_seq			**X**	

Table 1. Subset of reduction, simplification and elision rules

A quick comparison of these rules between varieties shows that IB01 and standard São Paulo share a total of eight rules, with IM01 and standard Lisbon sharing a total of seven. Shared rules between IM01 and standard São Paulo, and IB01 and standard Lisbon are fewer, consisting primarily of rules belonging to the 'do_r_4' subset. Those

rules that apply solely to IB01 and/or IM01 are also easily identified, and show in some small way the manner of innovation by which non-standard spoken forms are realized apart from corresponding standard varieties. This is further illustrated in Table 2, which shows the complete set of rules, organized by category, that were applied just to IB01 and/or IM01.

rule type	rule	IB01 (Belém)	IM01 (Maputo)
assimilation	**do_assimilated_s**	**X**	**X**
fricativization/affrication	**do_affricated_t_II**	**X**	**X**
homorganic nasal epenthesis	**do_homorganic_nasal**		**X**
lengthening	**do_aspirated_d_resyllabify**		**X**
lengthening	**do_long_d_resyllabify**		**X**
nasalization	**do_nasalize_I**	**X**	**X**
nasalization	**do_nasalize_II**		**X**
reduction/simplification/elision	**delete_final_i_resyllabify**		**X**
reduction/simplification/elision	**delete_initial_e_resyllabify**	**X**	
reduction/simplification/elision	**denasalize**	**X**	
reduction/simplification/elision	**do_e_h_resyllabify**		**X**
reduction/simplification/elision	**do_initial_I**		
reduction/simplification/elision	**do_schwa_II**	**X**	
reduction/simplification/elision	**simplify_affricate_seq**	**X**	
reduction/simplification/elision	**simplify_nasal_diphthong**	**X**	**X**
reduction/simplification/elision	**simplify_vowel_seq**	**X**	
rhotic differentiation	**do_alveolar_approximant**		**X**
rhotic differentiation	**do_glottal_fricative**		
rhotic differentiation	**do_velar_fricative_III**	**X**	
velarization	**do_velarized_l_resyllabify**		**X**
vowel opening	**do_e_E**	**X**	
vowel opening	**do_initial_E**		**X**

Table 2. Subset of the total rules applied just to IB01 and/or IM01

The 22 rules presented in Table 2 represent 36 percent of the total number of post-lexical rules currently described in LUPo for generating accent-specific transcriptions for 125 word forms. While it is assumed that this figure will change with the modelization of additional informants and the expansion of LUPo's master lexicon, one can observe for the current data set that more than one-third of the rules are

'innovative', or differ from the dominating varieties. Of course, it must be remembered that this proportion is skewed by the fact that intra-accent variability is not yet properly accounted for by LUPo. The point is rather one of illustrating LUPo's potential as means of teasing apart both common and innovative phenomena across spoken varieties.

In terms of findings from the literature regarding the Belém variety, nearly all of the phenomena described by authors such as de Carvalho (2000), Lopez (1979), Vieira (1983), Brandão & Cruz (2005), and Nina (1991) were observed for IB01, with the exception of the palatal lateral [ʎ] in syllable onsets preceding the high vowel [i] (Oliveira & Razky 2010), which was unattested in this informant's speech data. The much more scant phonetic observations contained in the literature concerning Mozambican Portuguese were observable, in part, for IM01, who is also a native speaker of the Bantu language, Kitonga. Indeed, the most compelling observations for this informant concern what are clearly contact traces of Kitonga, such as the realization by this speaker of a geminated voiced obstruent [dd], e.g. *esperança*[dd]*o*, along with insertion of an epenthetic schwa in some consonant clusters, e.g. *om*[ə]*nisciente* — findings which lend evidence to the tenets asserted in Gonçalves (2010).

4.2. Dialectometric comparisons

With dialectometric studies on the rise in recent decades, work by authors such as Heeringa et al. (2006), Nerbonne et al. (2008), Valls et al. (2010), and Kessler (1995), has resulted in the evaluation of new and old methodologies for conducting quantitative comparisons across dialects. The Levenshtein Distance (LD) algorithm offers one reasonably well regarded methodology (e.g. Valls et al 2010) for calculating the phonetic distance across strings. This is a remarkably simple algorithm, whereby two strings of phonetic segments (representing a single lexical item) are aligned, and the cost of generating one from the other is tallied for each non-identical deletion, insertion, and substitution. The end result is a numeric score that represents the degree of phonetic distance separating two varieties.

Table 3 shows a set of aggregate LD scores for the different combinations of accent pairs. Scores were derived by comparing LUPo output strings for each accent

pair, which were then totaled and averaged across the 125 word forms currently stored in LUPo's master lexicon. Given findings in Heeringa et al. (2006) concerning optimal application of the LD for doing dialectometric comparisons, scores were not normalized for length.

std Lisbon ~ std São Paulo	2.96
std Lisbon ~ IB01 (Belém)	2.78
std Lisbon ~ IM01 (Maputo)	2.73
std. São Paulo ~ IB01 (Belém)	1.35
std. São Paulo ~ IM01 (Maputo)	2.83
IB01 (Belém) ~ IM01 (Maputo)	2.62

Table 3. Averaged aggregate LD scores for accent pairs

As with the rule comparisons described in the previous section, the results displayed in Table 3 do not account for intra-accent variability. Moreover, *unlike* the rule comparisons demonstrated in section 4.1, the averaged aggregate LD distances shown reveal nothing about the actual phenomena responsible for the relative distance separating two varieties. Nevertheless, these scores show some basic patterns reflecting our general assumptions about the relative proximity of the sample's two sub-national varieties, standard São Paulo and Belém, and the roughly comparable distances that appear to separate the sample's remaining accent pairs.

Use of the LD to compare the metaphonemic forms stored in LUPo's master lexicon with the corresponding phonetic output generated for each variety (Table 4) is similarly opaque in terms of yielding results that either support or challenge our assumptions about pan Lusophone phonetic variation. Here, it must be reemphasized that the metaphonemic forms stored in the master lexicon are not strictly phonological, as metaphonemic segments must be generalizable to all potential spoken varieties. Still, it is a matter of some curiosity that the scores for the standard Lisbon and São Paulo varieties are respectively so much larger and smaller than those for the idiolectal varieties IB01 and IM01. As LUPo is expanded, this and the previous mode of comparison may likely present still more curious patterns worthy of analysis.

master lexicon ~ std Lisbon	2.57
master lexicon ~ std. São Paulo	2.06
master lexicon ~ IB01 (Belém)	2.56
master lexicon ~ IM01 (Maputo)	2.29

Table 4. Averaged aggregate LD scores for input and output pairs

5. Conclusions and future work

The work of the LUPo project has been described concerning the development of an accent-independent lexicon and rule system for generating phonetic transcriptions for regional accents of Portuguese. An initial prototype of the online LUPo system was presented, along with a window into the phonetic segmental modeling of Luso-African idiolectal varieties from Belém (BR) and Maputo (MZ).

It has been shown that LUPo is designed to handle variability at the national and sub-national levels. This is achieved economically, through the sharing of rules across pluridimensional varieties, as demonstrated in the description of LUPo's regional accent hierarchy, while acknowledging those salient segmental features that are essential in distinguishing one variety from another, and which result in more "natural" transcriptions. It was also shown that LUPo's output data, along with the metaphonemic forms and rules that go into making the LUPo system, present a range of opportunities for analyzing the distance between varieties, with the comparison of both shared and innovative rules potentially offering a more informative mode of analysis. In general, LUPo is poised to provide linguists with a huge list of varying points and bundled phenomena — along with tangible data links — for testing notions of linguistic similarity and distance, and evaluating the pulling effect of different linguistic centers.

In this vein, the LUPo project seeks to contribute to the improvement of Portuguese language speech technologies by providing high-quality pronunciation lexica, derived from linguistic rules, and covering as many topolectal variants as possible. It is further anticipated that this work will have a positive impact on raising the

profile of non-standard, "digitally endangered" (Rusko et al. 2008) varieties of Portuguese, contributing in some small part towards their enhanced prestige, and the perception of these varieties as worthy of study in their own right.

Future work will involve expanding the master lexicon to a list of 1500 high-frequency words; further development of the Belém and Maputo models; and the expansion of LUPo to include non-standard accents from Luanda (Angola), Rio de Janeiro and São Paulo (Brazil), the island of Praia (Cape Verde), Macau (China), Dili (East Timor), Nampula (Mozambique), and Braga (Portugal). Efforts are also under way to develop and launch a free, online, searchable database for use by the research community to test the results of different speech processing systems, conduct empirical analyses across multiple Portuguese accents, and facilitate second (or foreign) language studies of Portuguese.

Acknowledgements

The authors gratefully acknowledge the full support of the Fundação para a Ciência e a Tecnologia (PTDC/CLE-LIN/100335/2008), and the cooperation of Dr. Susan Fitt, whose development of the original English Unisyn Lexicon is the inspiration for this work.

References

BAXTER, Alan N. (1992) "Portuguese as a pluricentric language", in Michael G. CLYNE (ed.), *Pluricentric Languages: Differing norms in different languages*, Berlin: Mouton de Gruyter, 11-44.

BOERSMA, Paul & David WEENINK (2010) "Praat: Doing Phonetics by Computer", computer program, vs. 5.1.43. Retrieved from http://www.praat.org 4 August 2010.

BRANDÃO, Silvia Figueiredo & Maria Luiza de Carvalho CRUZ (2005) "Um estudo contrastivo sobre as vogais médias pretônicas em falares do Amazonas e do Pará com base nos dados do ALAM e do ALISPA", in V. AGUILERA (ed.), *A geolinguística no Brasil: caminhos e perspectivas*, Londrina: Editora da Universidade Estadual de Londrina, 299-318.

CAGLIARI, Luiz Carlos (1981) *Elementos de fonética do Português Brasileiro*, Tese do Título de Livre Docente, Universidade Estadual de Campinas.

CARVALHO, Maria José (1991) "Aspectos sintáctico-semânticos dos verbos locativos no Português oral de Maputo", Lisbon: *ICALP Angolê – Artes e Letras*, 145-152.

CARVALHO, Rosana Siqueira de (2000) *Variação do /s/ pós-vocálico na fala de Belém. Dissertação de mestrado*, Belém: UFPA.

CHIMBUTANE, Feliciano (1998) "As estratégias resumptiva e cortadora na formação de orações relativas do Português de Moçambique", in Perpétua GONÇALVES (ed.), *Mudanças do Português em Moçambique: Acquisição e formato de estruturas de subordinação*, Maputo: Universidade Eduardo Mondlane e Livraria Universitária, 111-181.

CIA (2010) The World Factbook–Mozambique. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/mz.html 10 December 2010.

CONCEIÇÃO, Manuel da (1999) "A Brief Look at the Sociolinguistics of Ronga and Other Languages Spoken in Mozambique", in University of Washington *Working Papers in Linguistics*, 16, 9-30.

COUTO, Mia (1986) "Uma maneira moçambicana de contar histórias moçambicanas", from an interview in Gazeta de Artes e Letras, *Tempo*, 835.

DIAS, Hildizina (2009) "A norma padrão e as mudanças linguísticas na língua Portuguesa nos meios de communicação de massas em Moçambique", in Hildizina DIAS (ed.), *Português Moçambicano: Estudos e reflexões*, Maputo: Imprensa Universitária, 389-420.

DINIZ, Maria João (1988) "Análise de erros: uma experiência com alunos moçambicanos", in Direcção da Associação Portuguesa de Linguística, *Actas do 3.º Encontro da Associação Portuguesa de Linguística*, Lisbon: Associação Portuguesa de Linguística, 125-138.

FITT, Susan (2000) "Documentation and user guide to UNISYN Lexicon and post-lexical rules". Technical report. The Centre for Speech Technology Research, University of Edinburgh.

FITT, Susan & Stephen ISARD (1999) "Synthesis of Regional English Using a Keyword Lexicon", in *Proceedings of Eurospeech*, Budapest, Hungary, 823-826.

GONÇALVES, Perpétua (1986) "Análise de erros em construções de subordinação", *Limani*, 1, 11-23.

GONÇALVES, Perpétua (2010) *A génese do Português de Moçambique*, Lisbon: Imprensa Nacional-Casa da Moeda.

HEERINGA, Wilbert, Peter KLEIWEG, Charlotte GOOSKENS & John NERBONNE (2006) "Evaluation of String Distance Algorithms for Dialectology", in John NERBONNE & Erhard HINRICHS (eds.), *Linguistic Distances*. ACL/COLING, Sydney, Australia, 51-62.

IBGE - Instituto Brasileiro de Geografia e Estatística, "Sintese de Indicadores Socias 2008", Belém, Brasil: IBGE. Retrieved from http://www.censo2010.ibge.gov.br 14 December 2010.

ISSAK, Aíssa (1998) "Estruturas de complementação verbal do Português de Moçambique", in GONÇALVES, Perpétua (ed.), *Mudanças do Português em Moçambique: Acquisição e formato de estruturas de subordinação*, Maputo: Universidade Eduardo Mondlane e Livraria Universitária, 67-110.

KESSLER, Brett (1995) "Computational Dialectology in Irish Gaelic", in *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, San Francisco: Morgan Kaufmann Publishers, 60-66.

LEVENSHTEIN, Vladimir I. (1965) "Binary codes capable of correcting deletions, insertions and reversals" *Doklady Akademii Nauk SSSR*, 163, 845-848.

LOPES, Armando Jorges (1979) "Mozambican Portuguese Words and Expressions", a lexical survey commissioned by Longman, and included in the *Longman English Dictionary for Portuguese Speakers* (1980), Harlow: Longman ELT.

LOPES, Armando Jorges (1999) "The Language Situation in Mozambique", in Robert KAPLAN B. & Richard B. BALDAUF, Jr. (eds.), *Language planning in Malawi, Mozambique, and the Philippines*, Tonawanda, New York: Multilingual Matters Ltd., 86-132.

LOPEZ, Barbara Stroot (1979) *The Sound Pattern of Brazilian Portuguese: Cariocan Dialect*, California: University of California Ann Harbor, University Microfilms, International.

MACHUNGO, Inês (2000) *Neologisms in Mozambican Portuguese: A Morphosemantic Study. University of Ghana*, doctoral thesis.

MACIEL, Carla & Joaquina PASCOAL (2002) "Produção científica sobre o Português de Moçambique", in Armindo NGUNGA, Samima PATEL, Inocêncio PEREIRA & Aurélio SIMANGO (eds.), I*nvestigação em ciências sociais e humanas: Situação actual e perspectivas*, Maputo: Livraria Universitária/Universidade Eduardo Mondlane, 1-93.

MATEUS, Maria Helena & D'ANDRADE, Ernesto (2000) *The Phonology of Portuguese*, New York: Oxford University Press.

NEWITT, Malyn (2002) "Mozambique", in Patrick CHABAL (ed.), *A History of Postcolonial Lusophone Africa*, London: Hurst & Company, 185-235.

NERBONNE, John, Peter KLEIWEG, Franz MANNI & Wilbert HEERINGA (2008) "Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering", in PREISACH, Christine, Lars BURKHARDT, Hans SCHMIDT-THIEME & Reinhold DECKER (eds.), *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, Berlin: Springer, 674-654.

NINA, Terezinha (1991) *Aspectos da variação fonético-fonológica na fala de Belém*, Rio de Janeiro: UFRJ.

OLIVEIRA, Marilucia Barros de & Abdelhak RAZKY (2010) "Imagens preliminares da realização variável de /l/ em posição prevocálica do norte, nordeste e centro-oeste do Brasil", in Mª João MARÇALO et al. (eds.), *Língua Portuguesa: Ultrapassar Fronteiras, Juntar Cultura*, Évora: Universidade de Évora, 163-183.

RODRIGUES, Maria Celeste Matias (2003) *Lisboa e Braga: Fonologia e Variação*, Lisbon: Fundação Calouste Gulbenkian.

RUSKO, Milan, Sakhia DARJAA, Marián TRNKA, Viliam ZEMAN & Juraj GLOVŇA (2008) "Making speech technologies available in (Serviko) Romani language", in Petr SOJKA, Aleš HORÁK, Ivan KOPEČEK & Karel PALA (eds.), *Lecture Notes in Artificial Intelligence*, 5246, 501-508.

SCHERRE, Maria Marta Pereira & Alzira V. Tavares MACEDO (1991) "Variação e Mudança: O caso do S pós-vocálico", *Boletim da Associação Brasileira de Lingüística*, 11, 165-80.

SITOE, Bento & Armindo NGUNGA (2000) *Relatório do II seminário sobre a padronização da ortografia de línguas Moçambicanas*, Maputo: NELIMO – Centro de Estudos das Línguas Moçambicanas, Universidade Eduardo Mondlane.

TRUDGILL, Peter (1983) *On Dialect*, New York: Basil Blackwell.

VALLS, Esteve, John NERBONNE, Jelena PROKIC, Martijn WIELING, Esteve CLUA & Maria-Rosa LLORET (2010) "Applying the Levenshtein Distance to Catalan Dialects: A Brief Comparison of Two Dialectometric Approaches", to appear in *Verba*. Anuario Galego de Filoxía, 37. Retrieved from http://ub.academia.edu 12 September 2010.

VIEIRA, Maria de Nazaré da Cruz (1983) *Aspectos do falar paraense*, Belém: Universidade Federal do Pará.

WELLS, John C. 1982. *The Accents of English*, New York: Cambridge University Press.