Large Scale Data Based Linguistic Investigations Using Speech Technology Tools: the Case of Romanian

Ioana Vasilescu
LIMSI–CNRS
Rue John Von Neumann
91400 Orsay France
Email: ioana.vasilescu@limsi.fr

Camille Dutrey
LPP – Paris 3 University & CNRS
19 Rue des Bernardins
75005 Paris France
Email: camille.dutrey@limsi.fr

Lori Lamel
LIMSI–CNRS
Rue John Von Neumann
91400 Orsay France
Email: lori.lamel@limsi.fr

***Abstract*—In this paper a first sketch of the Romanian vowels as illustrated by a 7 hours corpus of broadcast news data initially developed for ASR purposes is drawn. Data confirm a seven monophtongs vocalic system. Data underline acoustic overlap and complementary distribution of the non low central vowels [ɨ] and [ʌ]. The current findings supports previous hypotheses built on laboratory data and encourages further investigations on large scale data.**

1. Introduction

Last decades advances in computation and data storage have triggered a true revolution in the field of humanities with the ability to analyze large scale corpora that are thousands of times larger than the data sets of the past [31]. Among the humanities fields, linguistics and more specifically phonetic studies have seen a substantial benefit as the digital progress allowed both gathering large scale corpora of natural speech and the integration of speech technology in phonetic investigations [27].

This new way of conducting phonetic research has been made possible in particular through the tools issued from the automatic speech recognition (ASR) research, which made available large amount of aligned data [2]. It is important to note that for many years, researchers in phonetics have used small data sets recorded with a reduced number of speakers in laboratory conditions. This approach has been adopted for both practical (material) and scientific reasons. Firstly, it was taking several decades until both recording and storage tools became available and easy-to-use by "humanists" and until speech technology has been "diverted" from its normal objective for linguistic investigation purposes [28]. Secondly, the reduced amount of data allows a more accurate control of the available phenomena to study and avoid variation. For example, in such data the lexical coverage is often circumscribed to portray the targeted phenomenon. Indeed, the main objection to raise to the approach which makes use of large scale corpora is that they contains to much variation meaning more factors to control and evaluate and possibly an under-representation of the target aspect to study [30]. As a consequence, even today exploring linguistic hypotheses through statistical modeling of large scale corpora remains a challenging step for many linguists. However, the process is in motion and with the advent of significant volume of digitized speech and the advances in automatic alignment techniques, it is now possible to combine the benefits of working on big corpora with the fine detail and naturalness of speech recorded in "clean" (i.e. laboratory) conditions [29].

The key advantage of using this methodology is that it allows to study aspects of *variation* in speech. Speech is known to be highly variable in time and space, many factors contributing to this variability: speakers, situations, recording conditions. Variation concerns all linguistic levels, from segmental realizations to syntax and discourse. Large corpora have the advantage to provide instances of variation that can be statistically modeled. Understanding and modeling elements of variation may be beneficial for both linguistics and speech technology. With reference to speech technology, while today's best ASR speech models are more and more efficient, they have not yet reached the status of being able to perfectly take into account all observed acoustic variation [32]. As for linguistic domains, describing and quantifying variation may help understanding processes that lead to language evolution [33].

In this framework, *less-resourced languages* represent a specific challenge which concerns both speech technology and corpus linguistic studies. Indeed, the today world's globalization engender the necessity to bridge the numerical gap between technologically privileged countries with the developing world. Linguistic studies are directly concerned by this progress as less-resourced languages often lack of studies driven by the spoken version of the language. Romanian is described as a less-resourced language [14]. Romania joined the European Union (EU) in 2007 and benefits since then from the increasing effort to extend language technologies to less studied European languages. By extension, corpus linguistics research on Romanian takes advantage of the development of automatic tools with afferent digitized corpora.

In this paper we provide a preliminary analysis of Romanian vocalic system based on a large corpus initially developed to build an ASR system. The paper is organized as following: Section 2 sketches a description of Romanian;

Section 3 describes an ASR system built for Romanian; Section 4 put forward the usefulness of ASR errors as cues for variation; Section 5 describes preliminary linguistic investigations dedicated to Romanian inflection; Section 6 is dedicated to the investigation of Romanian vocalic system and conclusions and perspectives are discussed in Section 7.

2. Romanian language: brief overview

Romanian belongs to the new EU languages poorly represented in the speech technology world whereas the presence of their speakers across enlarged Europe constitutes a real challenge for such technologies. For instance, according to [14], ASR is one of the less represented voice-driven technology dedicated to Romanian language. Spoken by over 29 million speakers around the world, Romanian is mother tongue for 25 millions of speakers and official language of two countries: Romania and Republic of Moldavia [14].

Romanian is a Latin language, from the Oriental branch. Romanian was isolated, geographically and politically, from the other Romance languages during centuries, as a result preserving Latin features lost in the other Romance languages. Romanian is based on the late Vulgar Latin, being among the last territories conquered by the Roman Empire. However, the great part of the fundamental vocabulary has Latin origin (about 60%) as well as the morpho-syntax. Romanian is surrounded by Slavic languages, which determined numerous borrowings, in particular at the lexical level. The Slavic influence has been reinforced by a long use of the Cyrillic alphabet (introduced in Romania with the oriental Christian religion and adapted to write the Romanian language).

After the 18th century, Romanians, proud of their Latin origins borrow many "cultismos" from other Romance languages and in particular from French and Italian. In the history of the Romanian language a "re-latinization" of the language occurs [15], [16]. Finally, political, economic and social aspects in the Romanian history explain other Eastern European influences: Turkish, Greek, German, Hungarian etc. Today, the English influence became particularly important at the vocabulary level. Today's Romanian may be described as a Latin language (phonetic, morpho-syntactic and lexical levels), with strong Slavic influences (phonetic and lexical levels) but also with contemporaneous Romance and English elements (lexical level).

3. ASR system for Romanian

A Romanian ASR system was built within the Quaero program[1]. The Romanian system development is lightly supervised as no detailed annotations are available for the training data [11]. Studies on large vocabulary continuous speech recognition systems for Romanian are lacking, however attempts to build Speech-to-text (STT) systems on limited data have been made recently [3], [6].

1. http://www.quaero.org/

A corpus of 3.5 hours of speech with exact transcription was used as development (*dev*) data and more than 400 hours of audio data were employed to train the system. Selected data consist in various Broadcast News shows, from read speech to more spontaneous interactions (Euranet, RFI Journal, RRA (Radio Romania Actualitati), Antena 3). Recordings are of both male and female speakers and the number of speakers per source vary from 3 (Euranet) to 24 (Antena 3). For this preliminary work on Romanian, attention has been payed to avoid sources with significant number of overlapping contexts, foreign or regional accents and noisy background.

The phone set used for the Romanian system contains 33 phones: 20 consonants, 3 semi-vowels, 7 vowels and 3 special symbols (see Table 1). The correspondence between letters and phones is almost one-to-one. About twenty rules were used to transform letters into phones. Foreign words were manually phonetized.

IPA	Ex. Romanian	IPA	Ex. Romanian
p	**p**as	b	**b**an
t	**t**are	d	**d**acă
k	**c**al	g	**g**ol
m	**m**ic	n	**n**or
f	**f**oc	v	**v**al
s	**s**are	z	**z**id
h	**h**orn	t͡s	**ţ**ara
r	**r**epede	l	**l**ung
ʃ	**ş**arpe	ʒ	**j**ar
t͡ʃ	**c**er	d͡ʒ	**g**er
a	**a**pa	e	**e**rou
i	**i**nsula	o	**o**ra
u	**u**da	ə	ud**ă**
ɨ	**î**nspre		
o̯a	f**oa**rte	j	**ia**pa
e̯a	m**ea**		
–	silence	–	breath
–	filler		

Table 1. PHONES USED IN THE ROMANIAN ASR SYSTEM.

The system architecture is described in [20]. The Word Error Rate (WER) is of 17.1% on the 3.5 hours of the *dev* data and of 19.9% on the evaluation conducted in 2012 within the Quaero project. More specifically, WER ranges from 8.3% to 23.5%, with higher error rates on the more spontaneous sources.

4. ASR errors as cues for variation

Previous studies underlined that the analysis of ASR errors may provide cues about the potential ambiguities of a language [2]. Transcription errors may highlight speech regions which are challenging for the ASR system. Such spoken regions may correspond either to intrinsic ambiguities of the language or to some type of intra- and/or inter-speaker variation still problematic for ASR modeling [13]. With regard to linguistic analysis, the errors may be indicators of local variation and may help in assessing if the

observed variation occurs randomly or tends to generalize to sound changes [17].

In [20] a description of the most frequent ASR errors was carried out. The analysis pointed out that confusions usually concern the verbal conjugation and the nominal declension. Inherited from Latin, conjugation and declension marks are mainly word final. Such affixes may be less carefully articulated and then subject to confusions (e.g. *survola* "he was flying" instead of *survolau* "they were flying"). Among the pre-word elements, verbal subjunctive (*să* [sə]) and reflexive (*s(e)* [se]) marks behave as auxiliary elements: short and acoustically poor, they may be deleted in the connected speech (e.g. **scoată* instead of *să scoată* "to extract").

5. Preliminary studies on Romanian inflection

To the best of our knowledge, using automatically segmented and annotated data to investigate linguistic aspects is an innovative approach for Romanian. Indeed, Romanian lacks of recent linguistic descriptions based on the spoken language, as most of the studies being driven by its written version. As for the phonetic studies, up to now studies have been based on data sets collected in laboratory conditions [21]. Consequently the intra- and inter-speaker variation marks in the connected speech and more general the phonetic variation in the contemporaneous Romanian have not been studied. From the linguistic point of view, such studies may increase the knowledge of the phenomena which contribute to the evolution of a language. From the ASR point of view, accounting for variation marks may contribute to a better pronunciation variant modeling.

As a preliminary study, two phenomena related to the morpho-phonology of Romanian and responsible for automatic transcription errors have been investigated [20], [22]:

(i) the masculine definite article -l has a variable realization (*sistemu* or *sistemul* "the system") and
(ii) final C palatalization marking inflections (*plop–plopi* [plop^{j}] poplar(s); *sap–sapi* [sap^{j}] "I/you dig"); *ban–bani*[ban^{j}] "money(s)" is only subtly audible, possibly due to devoicing of the palatal articulation.

Both are word-final phenomena thus susceptible to deletion in connected speech. This preliminary investigation pointed out that the presence of the masculine definite article seems to be strongly linked to type of speech: the more spontaneous is the recording (talk shows, debates) the less carefully the definite article pronounced. As for the final palatalization, the results underlined that the phenomenon covers various contextual realizations that goes from the absence of the final vocoid (45.5% of the occurrences) to truly palatalized consonants as in Russian, a language with phonemic opposition between plain and palatalized consonants (20.2% of the data). Final palatalization with Romanian consonant model corresponds to 32.3% of the occurrences.

6. Corpus-based study of Romanian vowels

ASR opened the way for linguistic investigations based on very large-scale spoken data. Spoken data bases developed within the ASR framework are enriched with corresponding orthographic transcriptions. Thereafter, the acoustic model training process generates segmentation into words and on a sub-word level, into phone segments. Beyond enabling the development of vocal technologies, those data are a new precious material for linguistic studies: the ASR systems may be used as a tool to highlight linguistic variation and to determine whether an observed phenomenon occurs randomly or fallows regular patterns [1].

In this study the ASR system for Romanian is used for speech alignment in order to extract acoustic and prosodic parameters as features to describe the Romanian vowels.

6.1. Description of the Romanian vowels

Romanian vocalic system exhibits seven monophthongs that may appear in both stressed and unstressed position, in open or closed syllables, in lexical roots of any length and in affixes. In addition to the seven vocalic phonemes, phonemic inventory of Romanian includes also two phonemic diphthongs: /o͡a/ and /e͡a/. Vowels may appear adjacent to any consonant. Within phonological forms, the root is typically followed by one or more morphological suffixes that place restrictions on the vowels that actually appear in word-final position [5], [24].

In [21] Romanian vowels are investigated through their acoustic (distribution in the vocalic space) and prosodic (duration) properties in a corpus obtained in laboratory conditions (i.e. vowels in target words recorded in carrier sentences). Results confirm a seven vowels vocalic space. They also point out that Romanian monophthongs are realized similarly in unstressed versus stressed conditions. Concerning the timbre of the vowels, a new transcription is proposed for the mid central vowel as [ʌ] instead of [ə].

6.2. Data and methodology

The analysis described below is based on *dev* and *eval* data set up to build the ASR system described in [20]. Table 2 sums up corpus specificities. Corpus can be described as portraying the standard version of the language, based on the southern dialect. The broadcast sources are those listed in Section 3. A forced alignment of the manual reference is realized with the mentioned ASR system. Afterwards, classical acoustic and prosodic parameters (F1, F2, F3, duration, fundamental frequency) are automatically extracted with Praat [25], following the methodology described in [23]. The corpus contains more than 56k words resulting in 300,174 phonemes, vowels and consonants. Vowels represent about 42% of the data, that is 125,501 vocalic tokens.

Two filtering procedures are applied. The first filtering procedure (*Filter 1*) is aimed to avoid vocalic outliers and it is based on physical parameters. The second filtering procedure is based on contextual considerations (*Filter 2*).

# Recording	22
Total duration	7h10
# Words	56 296
# Distinct words	8 683
# Speakers	95

Table 2. ROMANIAN CORPUS OVERVIEW

- *Filter 1*: Vocalic outliers correspond to measurement errors due generally to short and devoiced items. Vowels meeting the following criteria have been kept for further investigations: (i) voicing rate above 70% ; (ii) duration above 0.07 seconds (for female speakers) or 0.09 seconds (male speakers), since male speakers present more variability. After this filter procedure 34,557 vocalic tokens have been retained for analysis (27.5% of the total number of the vocalic segments).
- *Filter 2*: Romanian vowels combine to result frequently in various more-than-one-vowel sequences with different phonetic/phonemic status: diphthongs, glide plus vowels sequences, triphthongs, hiatus [26]. In addition to historically motivated vocalic sequences, morpho-syntactic processes are responsible for increasing dramatically the number of such items. In connected speech the true realization of the vowels within such structures is a key issue, as contraction phenomena which are spontaneous speech proper may affect the canonical timbre or the realization itself of the vowels. In order to avoid miss-interpretations of vocalic status and timbre due to such phenomena, in this preliminary study we kept only monophthongs appearing in consonant-vowel-consonant (CVC) contexts.

Further analysis is based on the 17,795 vocalic segments which remains after the implementation of the two filters (14.2% of the total number of vocalic segments).

6.3. Results

The following paragraphs are dedicated to the description of the vocalic system of Romanian through its distribution in the vocalic space, the frequency of the vocalic segments and the behaviour of the two central vowels. For the mid open central vowel we adopt the IPA transcription [ʌ] as in [21] which avoids confusion with the phonological implications of an encoding as *schwa*.

6.3.1. Romanian vowels distribution. Figure 1 shows the frequency of the vocalic segments in the corpus. Central open and front vowels exhibit some differences possibly due to the lexical coverage of the corpora, however results are globally consistent with [21], in particular with regard to the very low representation of the central vowels [ɨ] and [ʌ].

As for the place occupied by the Romanian vowels in the acoustic space, Figure 2 shows the dispersion of the 17,795 vocalic tokens in the F1/F2 space divided as function of

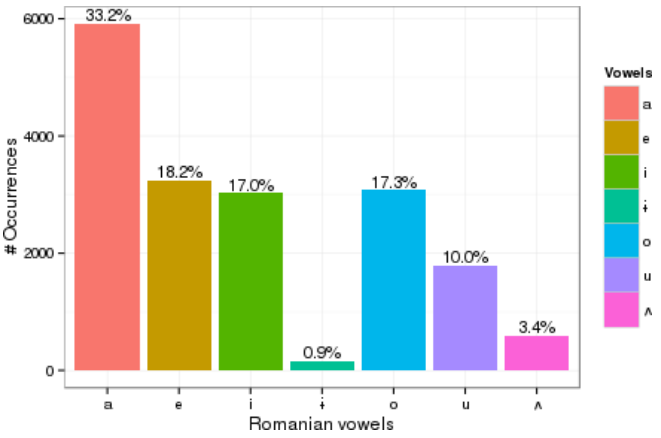

Figure 1. Frequency of the Romanian vowels in CVC contexts after *Filter 1* procedure.

the speaker gender. Romanian vowels fill a V-shape space, although the overlap between vowels being significant (in particular for central vowels) and requesting further filtering work. When considering the mean value per vowel type (Figure 3) one can notice the proximity between [ɨ] and [ʌ].

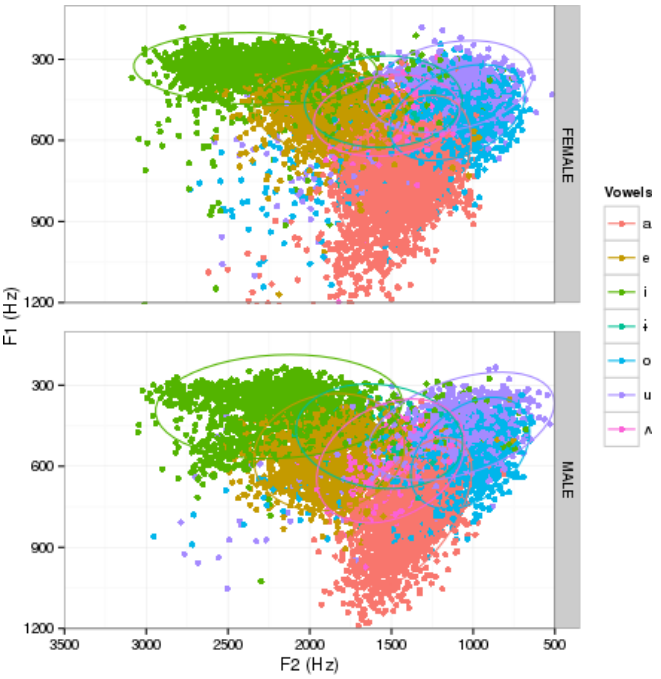

Figure 2. Romanian vowels distribution in the F1/F2 space for male and female speakers, all vocalic segments of the corpus with ellipses.

6.3.2. Timbre and status of the Romanian non-low central vowels. This paragraph is dedicated to the phonetic specificity of the non-low central vowels [ɨ] and [ʌ]. Among the seven vowels of the system, the two items have a particular status in Romanian: historically both are innovations in the Romanian vocalic system (compared with Latin

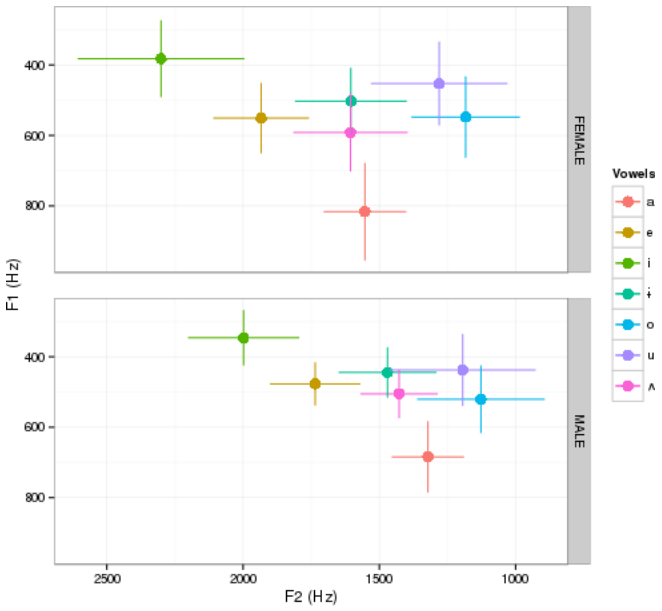

Figure 3. Romanian vowels distribution in the F1/F2 space for male and female speakers, mean and standard deviation of the vocalic segments of the corpus.

and other Romance languages) and spread in the language through both phonetic internal processes and foreign lexical borrowings. They are the less frequents in the language and although acoustically distinct in laboratory speech [21], they highly overlap in our data. An in-depth analysis of the lexical contexts in which the two vowels occur provides elements to justify their phonetic specificity.

The vowel [ɨ] corresponds to 3.4% of the vocalic segments in the broadcast news data analyzed in this study. [ɨ] may be found in 3.2% of the vocabulary that is 275 different word form. [ʌ] is also poorly represented, being the less frequent vowel (0.9% of the data). It can be found in 755 word forms, that is 8.7% of the vocabulary.

Table 3 shows [ɨ] and [ʌ] distribution as function of the surrounding vowel (V) and consonant (C) phonemes. Contexts are illustrated by the most frequent word in the corpus. Their occurrence in the present data is in line with previous findings supporting a complementary distribution of the two phonemes in Romanian. They appear predominantly in different contexts: [ɨ] is followed by nasal consonants whereas [ʌ] is frequent in word final position (e.g. as mark of nominal declension). It may be inferred that the acoustic proximity between [ɨ] and [ʌ] does not lead to word confusions as the lexical intersection of the items containing the two vowels is minimal. The present preliminary considerations support the hypothesis of marginal phonemic contrast proposed in [21].

7. Conclusion

This paper provides a summary of previous investigations dedicated to morpho-phonetics of Romanian based on 3.5 hours broadcast news corpus developed for ASR purposes and draws a preliminary description of Romanian

context	/ɨ/		/ʌ/	
CVC	30.5%	România	26.6%	astăzi
VVC	0.3%	neîncadrabili	2.8%	2012
CVV	1.3%	mâine	1.8%	său
#VC	67.2%	în	0.4%	ăsta
#VV	0.7%	îi	0.1%	ăia
CV#	–	–	65.8%	să
VV#	–	–	2.5%	două
ALL	100%		100%	

Table 3. CENTRAL VOWELS /ɨ/ AND /ʌ/ DISTRIBUTION CONSIDERING THEIR CONTEXT OF APPEARANCE. EXAMPLES WITH THE MOST FREQUENT WORDS FOR EACH CASE.

vowels as illustrated by a large scale corpus (more than 7 hours) of aligned data. We underline that Romanian, a language still described as a *low-resourced* is underrepresented in both speech technology studies and corpus linguistics. However, making use of digitized and aligned data coming from recent research dedicated to ASR may be highly innovative for classical linguistic investigations.

This study is dedicated to Romanian vowels characteristics from both distributional and acoustic perspectives. Data confirm a seven monophthongs vocalic system. The distribution of the vowels in the data shows that the non low central vowels [ɨ] and [ʌ] represent a minority of segments compared to the other vocalic realizations. Their presence in the language is mainly borne by lexical prefixes in nasal consonant context (i.e. [ɨ]) and nominal and verbal declension (i.e. [ʌ]). The two segments shows an acoustic overlap and the finding confirms previous observations which underline their minimal phonemic contrast.

Further studies will be conducted to get more insight on the phonetics of Romanian in order to link the behaviour of Romanian sounds to both the morpho-syntax of the language and to on-going evolutionary processes.

References

[1] Adda-Decker, M. and Lamel, L. 1999. Pronunciation variants across system configuration, language and speaking style. Speech Communication - Special issue on modeling pronunciation variation for automatic speech recognition. 29(2-4):83-98.

[2] Adda-Decker, M. 2006. De la reconnaissance automatique de la parole a l'analyse linguistique des corpus oraux. In *Proceedings of Journaux d'Etude sur la Parole*, France.

[3] Burileanu, C., Buzo, A., Petrea, C., Ghelmez-Hanes, D. and Cucu, H. 2010 Romanian spoken language resources and annotation for speaker independent spontaneous speech recognition. In *Proceedings of Conference on Digital Telecommunications* 7-10.

[4] Candea, M., Adda-Decker, M. and Lamel, L. 2013. Recent Evolution of Non Standard Consonantal Variants in French Broadcast News. In *Proceedings of Interspeech*, Lyon, France.

[5] Chitoran, I. 2002. A perception-production study of Romanian dipthongs and glide-vowel sequences. In *Journal of the International Phonetic Association*, 32:2003–2022.

[6] Cucu, H., Besacier, L., Burileanu, C. and Buzo, A. 2012 ASR Domain Adaptation Methods for Low-Resourced Languages: Application to Romanian Language. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, 1648-1652.

[7] Despres, J., Lamel, L., Gauvain, J.-L., Vieru, B., Woehrling, C., Le, V.B. and Oparin, I. 2013 The Vocapia Research ASR Systems for Evalita 2011. *Evaluation of Natural Language and Speech Tools for Italian* 286-294.

[8] Fousek, P., Lamel, L., and Gauvain, J.-L. 2008 Transcribing Broadcast Data using MLP Features. In *Proceedings of Interspeech 2008*, 1433-1436.

[9] Gauvain, J.-L., Lamel, L. and Adda, G. 2002 The LIMSI Broadcast News Transcription System. *Speech Communication,* 37(1-2):89-108.

[10] Lamel, L., Gauvain, J.-L and Adda, G. 2002 Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16:115-129.

[11] Lamel, L., and Vieru, B. 2010 Development of a speech-to-text transcription system for Finnish In *Proceedings of SLTU*, Malaysia, 62-67.

[12] Spinu, L., Vogel, I. and Bunnell, H.T. 2012. Palatalization in Romanian-Accoustic properties and perception. *Journal of Phonetics*, (40)1: 54-66.

[13] Vasilescu, I., Yahia, D., Snoeren, N., Adda-Decker, M. and Lamel, L. 2011. Crosslingual study of ASR errors: on the role of the context in human perception of nearhomophones. In Interspeech, Firenze, Italy.

[14] Trandabat, D., Irimia, E., Barbu Mititelu, V., Cristea, D. and Tufis, D. 2012. The Romanian Language in the Digital Age. In META-NET White Paper Studies, Springer.

[15] Hristea, Th. 1984., Sinteze de Limba Romana. Editura Albatros.

[16] Brancus, G. 1983., Vocabularul autohton al limbii romane (The Autochtone Vocabulary of Romanian Language). Editura Stiintifica si Tehnica (Scientific and Technical Publishing House).

[17] Ohala, J.J. 1996., The connection between sound change and connected speech processes. In Arbeitsberichte (AIPUK), Université Kiel, 201-206 .

[18] Dinu, L.P., Niculae, V. and Sulea, M., 2012., Dealing with the grey sheep of the Romanian gender system, the neuter. In Proceedings of COLING 2012, Mumbay, 119-124.

[19] Petrovici, E., 1956., Sistemul fonematic al limbii române (The phonematic system of Romanian). In Studii si Cercetari Lingvistice (Linguistic studies and Research), VII, f. 1-2, 7-20.

[20] Vasilescu, I., Vieru, B. and Lamel, L., 2014., Exploring Pronunciation Variants for Romanian Speech-to-Text Transcriptions. In Proceedings of SLTU, St. Petersburg.

[21] Renwick,M. 2014., The Phonetics and Phonology of Contrast : The Case of the Romanian Vowel System. Berlin: De Gruyter Mouton.

[22] Chitoran, I., Vasilescu, I., Vieru, B. and Lamel, L., 2014., Analyzing linguistic variation in a Romanian speech corpus through ASR errors. In LARP - Laboratory Approaches to Romance Phonology VII, Aix-en-Provence, France, Sept 3-5 2014.

[23] Gendrot, C. and Adda-Decker, M. 2005. Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In Proc. of Eurospeech, Lisbon, 2453-2456.

[24] Chitoran, I., 2002. The phonology of Romanian : a Constraint-Based Approach. In Studies in Generative Grammar 56 Berlin & New York, Mouton de Gruyter.

[25] Boersma, P. and Weenink, D. 1999. Praat, a system for doing phonetics by computer. Institute of Phonetic Sciences of the University of Amsterdam.

[26] Niculescu, O., Forthcoming. Hiatul intern si extern în limba română: O analiza acustica [Internal and external hiatus in Contemporary Standard Romanian: An acoustic analysis]. University of Bucharest PhD dissertation.

[27] 2011. New Tools and Methods for Very-Large-Scale Phonetics Research Workshop - VLSP-2011. Proceedings, University of Philadelphia.

[28] Lamel, L. 2015. Language Diversity: Speech Processing In A Multi-Lingual Context. Keynote speech, Interspeech 2014, Singapore.

[29] Coleman, J., Liberman, M., Kochanski, G., Burnard, L. and Joahong, Y. 2011. Mining a Year of Speech. Proceedings of New Tools and Methods for Very-Large-Scale Phonetics Research Workshop - VLSP-2011, University of Philadelphia.

[30] Kochanski, G.P., Shih, C. and Shosted, R. 2011. Should Corpora be Big, Rich, or Dense?. Proceedings of New Tools and Methods for Very-Large-Scale Phonetics Research Workshop - VLSP-2011, University of Philadelphia.

[31] Ide, N. 2004. Preparation and analysis of linguistic corpora. In A Companion to digital Humanities, Blackwell Publishing Ltd.

[32] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. and Wellekens, C. 2007. Automatic speech recognition and speech variability: a review. In Speech Communication, 49 (10-11), 763-786.

[33] Ohala, J.J. 1996. The connection between sound change and connected speech processes. Arbeitsberichte (AIPUK 31) Universität Kiel. 201-206.