Natural Language Engineering (2022), 1–26
doi:10.1017/S1351324922000419

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

RoLEX: The development of an extended Romanian lexical dataset and its evaluation at predicting concurrent lexical information

Beáta Lőrincz[1,*], Elena Irimia[2], Adriana Stan[3] and Verginica Barbu Mititelu[2]

[1]Babeş-Bolyai University, Cluj-Napoca, Romania, [2]Research Institute for Artificial Intelligence 'Mihai Dragănescu', Romanian Academy, Bucharest, Romania, and [3]Technical University of Cluj-Napoca, Cluj-Napoca, Romania
*Corresponding author. E-mail: beata.lorincz@ubbcluj.ro

(Received 9 December 2021; revised 19 July 2022; accepted 25 July 2022)

Abstract

In this article, we introduce an extended, freely available resource for the Romanian language, named RoLEX. The dataset was developed mainly for speech processing applications, yet its applicability extends beyond this domain. RoLEX includes over 330,000 curated entries with information regarding lemma, morphosyntactic description, syllabification, lexical stress and phonemic transcription. The process of selecting the list of word entries and semi-automatically annotating the complete lexical information associated with each of the entries is thoroughly described.

The dataset's inherent knowledge is then evaluated in a task of concurrent prediction of syllabification, lexical stress marking and phonemic transcription. The evaluation looked into several dataset design factors, such as the minimum viable number of entries for correct prediction, the optimisation of the minimum number of required entries through expert selection and the augmentation of the input with morphosyntactic information, as well as the influence of each task in the overall accuracy. The best results were obtained when the orthographic form of the entries was augmented with the complete morphosyntactic tags. A word error rate of 3.08% and a character error rate of 1.08% were obtained this way. We show that using a carefully selected subset of entries for training can result in a similar performance to the performance obtained by a larger set of randomly selected entries (twice as many). In terms of prediction complexity, the lexical stress marking posed most problems and accounts for around 60% of the errors in the predicted sequence.

Keywords: Lexical dataset; Romanian; Transformer; Concurrent lexical prediction

1. Introduction

Natural language processing (NLP) is increasingly present in all human-computer interaction applications. Personal assistants, machine translation engines, chat bots or speech synthesis and recognition systems enable a more immersive virtual experience. Yet all these applications rely on the availability of high-quality language resources, as well as accurate automated knowledge processing and extraction tools. The lack of any of these items hinders the development of state-of-the-art applications in the target language or language group.

The fundamental text processing tasks refer to basic lexical annotations of an orthographic transcript. The annotations commonly include lemmatisation, part-of-speech (POS) tagging and phonemic transcription. However, it is also common to require syllabification, lexical stress

Beáta Lőrincz, Elena Irimia, Adriana Stan, and Verginica Barbu Mititelu contributed equally.

Table 1. An example entry from the RoLEX dataset and the manner in which the lexical information was validated. The example entry is for the word *iarna*, the equivalent of the definite form of the English noun *'winter'*

Lexical information	Example	Validation method
Orthographic form	iarna	manual
Lemma	iarnă	manual
POS tag	Ncfsry	manual
Syllables	iar.na	automatic/manual
Stress	i'arna	automatic/manual
Phonemic transcription	j a r n a	automatic/manual

marking or complete morphosyntactic descriptors (MSD). Some of the applications that can benefit from the extended list of annotations include language learning interfaces, machine translation tools or most prominently speech-based applications, and especially text-to-speech synthesis (TTS) systems. For example, lemma and morphosyntactic information can help a machine translation system distinguish between homographs in a specific context. Exposing the correct phonemic transcription, lexical stress and syllabification sequence of a word can speed up the learning process of a foreign language. Phonemic transcription is also essential in speech recognition systems, where the models generally learn representations of the speech signal at phone-level (Zeineldeen *et al.* 2020). For TTS systems, the complete lexical annotation of the orthographic transcript is essential, and many recent studies augment the text input with this annotation and, as a result, enhance the naturalness and adequacy of the output speech (Peiró-Lilja and Farrús 2020; Taylor and Richmond 2020).

In this article, we describe the design and development of a large lexical dataset for Romanian which includes all the information enumerated above and the evaluation of the dataset's usability in predicting different lexical tasks. The main **contributions** of our paper can be summarised as follows:

[C1] We introduce RoLEX,[a] the largest freely available lexical dataset for Romanian with over 330,000 tokens. It includes information about lemma, POS, syllabification, lexical stress and phonemic transcription; **[C2]** we thoroughly describe the process of:

(1) selecting the words in the dataset based on a speech corpus,
(2) annotating them automatically with reliable *lemma* and POS information and partially reliable *syllabification, lexical stress marking* and *phonemic transcription* and
(3) validating, both automatically and manually, an important part of the entries: this was not an entry-by-entry validation, which would have implied an extensive manual work effort that we could not afford; instead, error patterns and entries with high error probability were automatically identified and manually or automatically corrected.

An overview of the information contained in RoLEX and the manner in which it was obtained and validated is presented in Table 1.

[C3] We analyse the accuracy of advanced neural network architectures in a task of concurrently predicting the *syllabification, lexical stress marking* and *phonemic transcription* from the context-free orthographic form of a word or from the orthographic form plus additional POS or MSD word tagging.

[a]The dataset can be found at the following URLs: www.racai.ro/p/reterom/results.html, https://github.com/adrianastan/rolex.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

2. Related work

2.1. The Romanian language and Romanian lexical datasets

Romanian is an Indo-European Romance language with a rich history of Slavic, German, Turkish and Hungarian influences. The Romance origin lends the highly inflected verb forms for person, number, tense, mood, and voice, while a large number of Slavic loan words influence its phonology.

With respect to the language particularities and their automated learning and prediction, Stan and Giurgiu (2018) acknowledge the regularities of the Romanian language that allow for implementing processing rules, but also enumerate the ambiguities that cannot be dealt with by these rules. For example, Romanian has 7 base syllabification rules (DOOM 2005), but there are several exceptions. Some are more general, like the hiatus-diphthong ambiguities or the different surfacing of the letter 'i' as a vocalic or non-vocalic element (Dinu, Niculae, and Sulea 2013). Others are rather particular, like the ones generated by foreign or compound words.

The Romanian phonetic spelling is generally a direct mapping from the orthographic form. Some exceptions are the two different sounds associated with letter 'x' (/cs/ and/gz/), or the several groups of letters (e.g., 'ce', 'ci', 'ge', 'gi', 'che', 'chi', 'ghe', 'ghi') that correspond to either a sound or two depending on the right-hand side phonetic neighbourhood of these groups. At the lexical level, non-homophone homographs are encountered. For example, the Romanian word 'hainǎ' can be pronounced[b] as/h a j n @/ (syl. hai.nǎ, stress: h'ainǎ, En. *coat*) or as/h a i n @/ (syl. ha.i.nǎ, stress: ha'inǎ, En. *wicked*). This example also illustrates the hiatus-diphthong ambiguity that affects syllabification.

Even though it serves to distinguish between different lemmas or between different forms of the same lemma, unlike other Romance languages such as Italian or Spanish, lexical stress is not graphically marked in written Romanian.

Although it has a relatively large number of native speakers (around 25 million), Romanian is still considered a low-resourced language in terms of digital resources and NLP tools (Trandabat *et al.* 2012). In the recent initiative called European Language Grid (Rehm *et al.* 2020), Romanian continues to be under-represented (with only 183 resources, tools and services) as compared to English (3039), Spanish (789), French (809) or German (934).[c]

The available lexical resources include RoSyllabiDict, NaviRo and MaRePhor. RoSyllabiDict (Barbu 2008) is a dictionary that contains syllabification and stress marking information for 525,534 word forms, corresponding to 65,000 lemmas. The resource was created by implementing the standard set of Romanian syllabification rules, using an inventory of Romanian diphthongs and triphthongs and the partial syllabification information provided in the database of the online Romanian Explicative Dictionary – Dexonline.[d] Dexonline was also the source for stress marking in RoSyllabiDict. The authors maintain that they performed partial validation of their resource at syllabification level. NaviRo (Domokos, Buza, and Toderean 2012) consists of more than 100,000 words extracted from Dexonline and transcribed in their phonemic form using an artificial neural network-based method starting from a seed set of manually transcribed entries. The authors report that they performed a manual check of the final dataset, but also mention that errors can still be found. MaRePhor (Toma *et al.* 2017) is a dictionary that provides phonemic transcription for 72,375 lemmas which make up the official list of the Romanian Scrabble Association. Therefore, this resource does not account for morphological variations. Two other resources, the Morphologic and Phonetic Dictionary of the Romanian Language (Diaconescu *et al.* 2015a) and the Phonetic Dictionary of Romanian Syllables (Diaconescu *et al.* 2015b), are only available as printed material.

[b] See Section 3.2 for the conventions of the phonemic transcription used for RoLEX.
[c] As per December 12, 2021, in the ELG Release 2: https://live.european-language-grid.eu/catalogue/
[d] www.dexonline.ro

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Given the scarcity of Romanian digital resources, as well as the disjoint information contained therein, we considered necessary to aggregate the different lexical information already available in some of the resources into a uniform dataset, with a principled bottom-up design and development. Additional requirements refer to the broad coverage of both morphological and lexical levels and the use of effective semi-automatic validation and correction steps.

2.1.1. Large lexicons for other languages

While for English the most known and used dataset is CMU Pronouncing Dictionary (CMUDict),[e] whose development as an open-source lexicon for speech recognition research started in the 90's, similar lexical datasets were gradually developed for other languages: PhonItalia (Goslin, Galluzzi, and Romani 2014) is a phonological lexicon for Italian that also includes syllabification and stress information, together with numerous lexical statistics for 120,000 word forms; for French, there is a phonetic lexicon (de Mareüil *et al.* 2000) comprising 310,332 word forms corresponding to 27,873 unique lemmas, and 10,000 proper names, with information about lemma, morphosyntactic description, automatically generated and partially corrected phonetic transcription; for European Portuguese, the web interface Procura-PALavras (P-PAL) (Soares *et al.* 2018) offers access to a lexical database based on a corpus of over 227 million words that contains very rich information (including morphosyntactic information, stress, syllabification and pronunciation) for around 208,000 word forms corresponding to approximately 53,000 lemmas; ArabLEX (Halpern 2022) is a very large lexicon covering over 530 million general vocabulary and proper noun words, with a variety of grammatical, morphological and phonological information, including stress and phonemic-phonetic transcription; LC-STAR German Phonetic lexicon[f] has 102,169 entries (55,507 common word entries extracted from a corpus of over 15 million words, 46,662 proper names and 6,763 words representing specific vocabulary for applications controlled by voice translated from English) with phonetic transcriptions given in SAMPA; the ILSP Psycholinguistic Resource (IPLR) (Protopapas *et al.* 2012) is a Greek lexical database with 217,000 entries comprising automatically generated information about syllabification, stress and phonetic transcription, while GreekLex2 (Kyparissiadis *et al.* 2017) is a lexical database that guarantees accurate syllabification, orthographic information predictive of stress and phonetic information for 35,000 words.

2.2. Lexical information prediction

To the best of our knowledge, the concurrent prediction and evaluation of all three lexical tasks is performed only in van Esch, Chua, and Rao (2016) and Lőrincz (2020). van Esch *et al.* (2016) uses an in-house dataset to improve the phonemic transcription and lexical stress marking by implicitly learning these tasks in a joint recurrent network-based sequence prediction. Lőrincz (2020) evaluates recurrent and convolutional (CNN) networks' performance in the concurrent prediction for English (41.04% WER) and Romanian (13.36% WER).

However, there are many studies which address the automatic annotation of each lexical task individually by employing different rule-based, traditional machine learning or deep learning approaches. Within these studies, the main focus language is English. For example, (Pearson *et al.* 2000) present decision tree-based methods for lexical stress and syllabification prediction. In (Webster 2004; Dou *et al.* 2009), decision tree and Support Vector Machine methods are evaluated for stress prediction and then used for grapheme to phoneme conversion modules in TTS systems. The phonemic transcription of English is also widely studied, and all levels of complexity algorithms were applied. The most recent approaches include neural sequence-to-sequence models, as

[e] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/
[f] https://catalogue.elra.info/en-us/repository/browse/ELRA-S0245/

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

described in (Yao and Zweig 2015; Milde, Schmidt, and Köhler 2017; Chae *et al.* 2018; Yolchuyeva, Németh, and Gyires-Tóth 2019a). The reported word error rates (WER) are between 20% and 25% for the CMUDict dataset. An encoder-decoder model with attention (Toshniwal and Livescu 2016) and a convolutional architecture combined with n-grams (Rao *et al.* 2015) achieve similar results when applied to the same dataset. Transformer-based architectures are proposed in Sun *et al.* (2019); Yolchuyeva, Németh, and Gyires-Tóth (2019b); Stan (2020) and slightly improve the error rates. Sun et al. (2019) report a WER around 20% obtained with a model enriched through knowledge distillation using unlabelled source words.

For Romanian, the three lexical tasks were also subject to a series of studies covering Marcus Contextual Grammars (Dinu 2003), rule-based methods (Toma and Munteanu 2009), decision trees and linear models (Cucu *et al.* 2014), cascaded sequential models (Ciobanu, Dinu, and Dinu 2014; Dinu *et al.* 2014) or neural network-based approaches (Boroș, Dumitrescu, and Pais 2018; Stan and Giurgiu 2018; Stan 2019, 2020). The reported WER for stress prediction is 2.36%, while the reported WERs for the phonemic transcription are between 1% and 3%. However, because the studies use different datasets, the error rates are not directly comparable to the results presented in this paper. We hope that with the availability of the RoLEX lexicon, future lexical information prediction tools will have a common reference point.

3. RoLEX development and validation

The development of RoLEX started within the ReTeRom project,[g] whose aim is to collect a large Romanian bimodal corpus, which can serve as training and testing material for improving available instruments for processing spoken and written Romanian. The corpus is a large collection of texts assembled from news articles, interviews on contemporary subjects, radio talk shows, tales and novels, and Wikipedia articles. The key characteristic of this corpus is its bimodality: it contains spoken Romanian language aligned with its written counterpart, either transcribed (in the case of interviews and talk shows) or originally written (in the rest of the cases). For the RoLEX development, only the written component was considered.

Being a corpus-based dataset (thus, a better representation of the language in use) makes it more appropriate for use in real-life applications. The corpus aggregated for RoLEX development contains transcriptions of the following speech corpora: the oral component of CoRoLa[h] (Barbu Mititelu, Tufiș, and Irimia 2018) (821,294 tokens), RSC (Georgescu *et al.* 2020) (590,190 tokens), SSC-train (1,262,030 tokens), SSC-eval (Georgescu, Cucu, and Burileanu 2017) (36,424 tokens) corpora, SWARA (Stan *et al.* 2017) (15,070 tokens) and MARA (Stan *et al.* 2021) (95,567 tokens). The quality of the starting corpus data varies from high-quality transcripts to documents that contain spelling and grammar errors or texts that lack punctuation, diacritics and capitalisation. Other subsets of the data are just lists of words or sub-sentential sequences. Some parts of the initial data collection were already processed: tokenised, lemmatised and POS-tagged with various degrees of correctness.

From the initial data collection, the first step in obtaining the lexical dataset was to extract a list of words containing correct contemporary Romanian words with no grammatical or spelling errors. The development of this dataset was based on a curated general lexicon of over 1.1 million entries of the Romanian language under development, called TBL,[i] containing lemma and morphosyntactic descriptor[j] information. The difference between TBL and RoLEX is the fact that the

[g] https://www.racai.ro/p/reterom/

[h] https://corola.racai.ro

[i] https://github.com/racai-ai/Rodna/blob/master/data/resources/tbl.wordform.ro

[j] The MSD follow the specifications developed for Romanian in the MULTEXT-EAST project: http://nl.ijs.si/ME/Vault/V5/msd/html/msd-ro.html

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

latter is derived from a set of contemporary texts, as opposed to just an exhaustive dictionary-like list of words as in TBL.

Two methodologies of lexicon extraction had to be adopted, depending on whether a reliable processed version of a document could be obtained or not. For the grammatically correct texts, the TEPROLIN web service (Ion 2018) was used to perform lexical segmentation (tokenisation), lemmatisation and POS-tagging.

The less accurate textual data could not be processed by automatic means since the tools are usually trained on correct grammatical texts and would, therefore, generate poor results on incorrect input data. In this situation, the lexical segmentation task is trivialised: the text was automatically tokenised at each blank space. The resulting tokens were checked against TBL for correctness. The contracted sequences, marked by a hyphen in Romanian, were treated separately. These sequences were segmented at the hyphen, which was successively attached to the different contraction components, generating two different possibilities for the segment tuples. The correct segmentation was identified by looking up the segments in TBL. For example, the sequence 'schimbându-şi' (En. *changing*-Cl.poss.refl.3) generated the tuples ('schimbându-', 'şi') and ('schimbându', '-şi'). The correct segmentation can be identified automatically by checking if TBL contains both terms of the segment tuple: for ('schimbându-', 'şi') we find that 'şi' is a word in TBL, but it has a different morphosyntactic annotation and role (namely the conjunction equivalent to the En. *and*) than the one intended in the sequence (reflexive/possessive pronominal clitic), and 'schimbându-' is not a correct Romanian word,[k] so the tuple is no longer considered as a possible segmentation; for ('schimbându', '-şi') both words occur in TBL, so this is the only correct segmentation of the sequence. Other sequences, like 'n-am' (En. *not-have*), both possible segmentations have the component words present in TBL ('n-', 'am', 'n' and '-am'); therefore, a manual examination is necessary to choose the right segmentation.

In a next step, TBL was used to identify all the entries linked to a specific form: if TBL contains the word form, all the corresponding (lemma, MSD) pairs associated with it and all the morphological variants of these lemmas are recovered and transferred to RoLEX. Treating the possible homonymy, which leads to POS and lemma ambiguities, was not a purpose at this step of generating RoLEX. Duplicated lexicon entries due to two methodologies used for different sub-corpora were searched for and eliminated. On the other hand, if a word form is not found in TBL, it is extracted in a separate list, to be manually validated and annotated.

We envisioned, from the very beginning, that the manual validation/correction work for the dataset will be time consuming and looked for strategies to make this work as efficient as possible. As described in Section 5, a more automatised and efficient technique for organising and reducing the manual correction effort can be employed, but at this point, the main solutions we found were (i) dividing the correction task by partitioning the dataset into parts with different risks and types of errors; (ii) automatising most of the correction tasks by means of linguistic rules.

Aside from the list of words which required complete or partial manual annotation, the automatically and semi-automatically generated lexical dataset was distributed to the correction team members for manual inspection, alongside instructions about the types of errors they needed to focus on. As all the annotators were expert linguists and the correction task was rather trivial, without ambiguities, we were not concerned with inter-annotator agreement and each data sample was distributed to only one annotator.

The automatic and semi-automatic annotation process, as well as the manual validation procedure with a focus on identified exceptions and rules are described in the next sections.

[k]The hyphen is used here to mark the absence of a sound from the clitic structure, namely the vowel *î*: the phonetically independent clitic *îşi* occurs in its dependent form *-şi*, which must be attached to an independent word, in this case the gerund. The hyphen also has the role of marking the pronunciation as one syllable of the clitic and the last syllable of the verb: 'du-şi'.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

3.1. Validation and annotation of lemma and morphological information

Because the largest part of the initial lexicon was obtained by querying TBL, the lemma and MSD for these entries were directly transferred to RoLEX. This solution also overcame the problem of incorrect morphosyntactic annotations found in the initial textual corpus.

For the new words, strategies for reducing manual work could also be applied in some cases, such as that of very productive morphological processes. A largely applied principle in lexicography is not to record exhaustive lists of words created by means of very productive derivation mechanisms, which are well mastered by a language's speaker. Words newly coined by means of these mechanisms are dealt with by recognising these productive rules and listing the components in the lexicon (e.g., very frequent prefixes, such as 're-' and the numerous verbal roots it attaches to). In our case, a specific type of new words is represented by a list of words formed by adding the prefix 'ne-' (En. *un-*) (even 'nemai-', in which the adverb 'mai' (En. *more*) is inserted between the prefix and the root) to gerund and participle forms of verbs; they were automatically dealt with by separating the prefix and the gerund suffix and looking up the roots in TBL, as they contain the lemma and MSD that also apply to the prefixed forms.

Only after automatising all the possible tasks, the remaining words were evaluated one by one and annotated with the corresponding lemma and MSD tag. Entries for their morphological variants were also created. Some frequently identified errors were typos (missing, extra or shifted letters), missing diacritics and lexical segmentation errors. 8,000 new entries (missing from TBL) were corrected/developed and integrated into RoLEX. They are also envisaged for the further extension of TBL.

3.2. Validation and correction for syllabification, stress marking and phonemic transcription

The rest of the lexical annotations–syllabification, lexical stress and phonemic transcription were partially obtained from the RoSyllabiDict and MaRePhor dictionaries. The entries not found in the two datasets were automatically annotated with the front-end tool developed in Stan *et al.* (2011). The tool, referred to as RoTTS, is used in text-to-speech synthesis systems and uses decision trees trained on a small in-house lexical dataset to predict each information individually.

The data were divided between entries coming from dictionaries and entries generated by the RoTTS tool. The starting hypothesis was that the two dictionaries primarily used for annotation (RoSyllabiDict and MaRePhor) were, as their authors claimed, partially validated before launching. Therefore, in theory, fewer errors for our dataset entries annotated based on these resources should have been encountered and the focus should have been more on the entries annotated with the RoTTS tool. In practice, MaRePhor has phonemic transcription only for words' lemmas, leaving their morphological variants to be annotated automatically. Although RoSyllabiDict offers syllabification and stress marking information for some morphological variants, the morphological paradigms are often incomplete. Also, both resources lack morphosyntactic information, which makes it impossible for RoTTS to correctly annotate ambiguous cases.

Some ambiguous entries are shown in the examples from Table 2. In Example 1, assigning the correct MSD annotation helps identify the right lemma of the word in focus, and therefore, it determines the syllabification, which in turn, according to rules concerning the phonemic transcription of the vowels and semi-vowels, determines the transcription. In Example 1.1, the initial 'i' is a vowel, while in 1.2 it is a semi-vowel part of the triphthong 'iei' always transcribed as/j e j/ (see Table 4 for more examples of hiatus/diphthong/triphthong occurrences of the same strings of letters). Example 2 is even more problematic: the two words share the POS and most of the morphological characteristics: type of noun (common), number (plural), case invariant, indefinite form; it is only the value of the gender attribute that distinguishes between the two words: masculine for Example 2.1 and feminine for Example 2.2. This type of specific ambiguity is rare in Romanian.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 2. Ambiguous entries for syllabification, stress marking and phonemic transcription

	Word	Lemma	En.	MSD	Syll	Stress	Phones
Example 1.							
1.	iei	ie	*blouse*	Ncfsoy	i.ei	’iei	/i e j/
2.	iei	lua	*(to) take*	Vmip2s	iei	i’ei	/j e j/
Example 2.							
1.	copii	copil	*child*	Ncmp-n	co.pii	c’opii	/k o p i j/
2.	copii	copie	*copy*	Ncfp-n	co.pii	cop’ii	/k o p i j/

For the data coming from the dictionaries, the correction stage targeted especially entries which had different lemmas and/or MSD descriptors associated with the same form. For the RoTTS generated annotations, many other types of possible errors were encountered and the correction benefited from further division of the task, as well as from the design of a set of lexical rules, as it will be described in the next section. Because the automatic validation/correction of phonemic transcription depends on applying rules on correct syllabification and stress marking information, the order in which the annotation levels are corrected is (i) syllabification; (ii) stress marking; (iii) phonemic transcription (except proper names and abbreviations, which are treated separately and corrected manually). For all the three annotation levels, in the first step a list of rules was derived from the data and used to automatically annotate or detect incorrect annotations of the entries. The result was then validated by the expert linguists.

3.2.1. Syllabification correction stage

In this step, we identified the situations which are likely to produce syllabification errors, as listed below:

- *words that contain syllables longer than four letters*: this is a rare case in the language: according to Dinu and Dinu (2006), 13% of the syllables in their corpus of 4,276 words contain at least 5 letters;
- *words that contain syllables with more than one vowel*: in Romanian, the letters ‘a’, ‘â’, ‘î’, ‘ă’ are always vowels, thus, syllables that have a combination of two of these letters are, therefore, incorrect;
- *words that contain letters that could represent either vowels or semi-vowels (see Table 3)*: this is the distinction between hiatus and diphthong or triphthong, that influences the transcription as a vowel or as a semi-vowel. The vowels involved in hiatus undergo vowel transcription, excepting the cases when they are involved in other diphthongs or triphthongs right near the hiatus (see examples in Table 4);
- *proper nouns and abbreviations*: the annotation for these words was automatically generated with RoTTS and contained many errors; some reasons for this are: foreign proper nouns usually preserve the pronunciation from their original language (which differ from the Romanian one in the case of many languages, for example English, German, French, Spanish, etc.), Romanian proper nouns may also have atypical pronunciations (e.g., some proper nouns are homographs of common nouns, but the two words are not homophones: ‘Curea’ stressed C’urea versus ‘curea’ (En. *belt*) stressed cure’a), the syllabification for the abbreviations is not well dealt with by RoTTS; 5,540 proper names and 373 abbreviations

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 3. Letters and letter groups that create ambiguities in the phonemic transcription. For the vocalic letters, we also note their vowel/semi-vowel phonemic value

Letter/group	Value	Phoneme	Example	Transcription
e	Vowel	/e/	eter, En. *ether*	/e t e r/
	Semi-vowel	/e_X/	neam, En. *nation*	/n e_X a m/
	Special pronunciation	/je/	este, En. *is*	/je s t e/
i	Vowel	/i/	vin, En. *wine*	/v i n/
	Semi-vowel	/j/	iar, En. *again*	/j a r/
	Whispered 'i'	/i_0/	pomi, En. *trees*	/p o m i_0/
o	Vowel	/o/	acolo, En. *there*	/a k o l o/
	Semi-vowel	/o_X/	soare, En. *sun*	/s o_X a r e/
u	Vowel	/u/	sur, En. *grey*	/s u r/
	Semi-vowel	/w/	sau, En. *or*	/s a w/
ce		/tS/	ceas, En. *clock*	/tS a s/
			cercei, En. *earrings*	/tS e r tS e j/
ci		/tS/	cine, En. *who*	/tS i n e/
			ciupi, En. *(to) pinch*	/tS u p i/
che		/k_j/	chema, En. *(to) call*	/k_j e m a/
			cheag, En. *clot*	/k_j a g/
chi		/k_j/	chip, En. *face*	/k_j i p/
			chiar, En. *even*	/k_j a r/
			rochii, En. *dresses*	/r o k_j i j/
ge		/gZ/	ager, En. *agile*	/a gZ e r/
			geană, En. *eyelash*	/gZ a n @/
gi		/gZ/	legifera, En. *(to) legislate*	/l e gZ i f e r a/
			magiun, En. *jam*	/m a gZ u n/
ghe		/g_j/	ghem, En. *(yarn) ball*	/g_j e m/
			lighean, En. *basin*	/l i g_j a n/
ghi		/g_j/	ghindă, En. *acorn*	/g_j i n d @/
			ghiară, En. *claw*	/g_j a r @/
k		/k/	karat, En. *karat*	/k a r a t/
		/k_j/	kilogram, En. *kilogram*	/k_j i l o g r a m/
q		/k/	Qatar	/k a t a r/
qu		/k_j/	Maquis	/m a k_j i s/
x		/ks/	exonera, En. *(to) exonerate*	/e ks o n e r a/
		/gz/	examen, En. *exam*	/e gz a m e n/

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 4. Examples for the use of vowel sequences in Romanian as hiatus or as diphthong or triphthong.

Seq	Hiatus	Diphthong/Triphthong	Type
ai	î.n**a.i**n.te	cân.tai	desc
	/1 n **a i** n t e/	/k 1 n t **a j**/	
	(En. *before*)	(En. *sing*, past cont., 2 sg.)	
au	**a.u**r	cân.t**au**	desc
	/**a u** r/	/k 1 n t **a w**/	
	(En. *gold*)	(En. *sing*, past, 3 pl.)	
ei	n**e.i**s.pră.vit	tei	desc
	/n **e i** s p r @ v i t/	/t **e j**/	
	(En. *unfinished*)	(En. *linden*)	
eu	n**e.u**.tru	l**eu**	desc
	n **e u** t r u/	/l **e w**/	
	(En. *neutral*)	(En. *lion*)	
ii	ş**ti.in.ţă**	co.p**ii**	desc
	/S t **i i** n ts @/	/k o p **i j**/	
	(En. *science*)	(En. *children*)	
oi	v**o.**in.ţă	bu.t**oi**	desc
	v **o i** n ts @/	/b u t **o j**/	
	(En. *will*)	(En. *barrel*)	
ou	bi.r**o.u**l	e.c**ou**	desc
	/b i r **o u** l/	/e k **o w**/	
	(En. *the desk*)	(En. *echo*)	
ui	în.gă.d**u.i**	în.gă.d**ui**	desc
	1 n g @ d **u i**/	/1 n g @ d **u j**/	
	(En. *to allow*, inf.)	(En. *to allow*, pres., 1 or 2 sg.)	
ăi	tr**ă.i**nd	c**ăi**	desc
	/t r **@ i** n d/	/k **@ j**/	
	(En. *to live*, ger.)	(En. *ways*)	
ău	c**ă.u**.tând	du.l**ău**	desc
	c **@ u** t 1 n d/	/d u l **@ w**/	
	(En. *to search*, ger.)	(En. *big dog*)	
âi	mâ.r**â.i**	c**âi**.ne	desc
	/m 1 r **1 i**/	/k **1 j** n e/	
	(En. *to grawl*)	(En. *dog*)	

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 4. Continued.

Seq	Hiatus	Diphthong/Triphthong	Type
âu	pâ.r**â.u**l	pâ.r**âu**	desc
	p 1 r **1 u** l/	/p 1 r **1 w**/	
	(En. *the stream*)	(En. *stream*)	
ea	r**e.a**l	r**ea**	asc
	/r **e a** l/	/r **e_X a**/	
	(En. *real*)	(En. *mean*, fem.)	
eo	ar.h**e.o**.log	vr**eo**	asc
	a r h **e o** l o g/	/v r **e_X o**/	
	(En. *archaeologist*)	(En. *some*)	
ia	spe.r**i.a**	p**ia**.tră	asc
	/s p e r **i a**/	p **j a** t r @/	
	(En. *to frighten*)	(En. *stone*)	
iu	b**i.u**.ni.voc	f**iu**	desc
	/b **i u** n i v o k/	/f **i w**/ (En. *son*)	
	(En. *two-way*)	**iu**.b'it	asc
		/**j u** b i t/ (En. *loved*)	
ie	sa.n**i.e**	f**ie**r	asc
	/s a n **i e**/	f **j e** r/	
	(En. *sleigh*)	(En. *iron*)	
io	b**i.o**.lo.gi.e	m**io**r.lă.i	asc
	b **i o** l o gZ i e/	/m **j o** r l @ i/	
	(En. *biology*)	(En. *to whine*, inf. or past 3 sg.)	
oa	c**o.a**.li.ți.e	**oa**.meni	asc
	/k **o a** l i ts i e/	/**o_X a** m e n i_0/	
	(En. *coalition*)	(En. *people*)	
ua	ac.t**u.a**l	zi.**ua**	asc
	/a k t **u a** l/	/z i **w a**/	
	(En. *current*)	(En. *the day*)	
uă	per.pe.t**u.ă**	do.**uă**	asc
	/p e r p e t **u @**/	/d o **w @**/	
	(En. *perpetual*, fem.)	(En. *two*, fem.)	
eai	a.gr**e.ai**	do.r**eai**	centred
	[a g r **e a j**/	/d o r **e_X a j**/	
	(En. *to agree*, past cont., 2 sg.)	(En. *to wish*, past cont., 2 sg.)	

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 4. Continued.

Seq	Hiatus	Diphthong/Triphthong	Type
eau	a.gr**e.au**	ce.r**eau**	centred
	/a g r **e a w**/	/tS e r **e_X a w**/	
	(En. *to agree*, past cont., 3 pl.)	(En. *to ask for*, past cont, 2 pl.)	
iai	scr**i.ai**	tă.**iai**	centred
	/s k r **i a j**/	/t @ **j a j**/	
	(En. *to write*, past cont., 2 sg.)	(En. *to cut*, past cont., 2 sg.)	
iau	scr**i.au**	tră.**iau**	centred
	/s k r **i a w**/	/t r @ **j a w**/	
	(En. *to write*, past cont., 3 pl.)	(En. *to live*, past cont, 3 pl.)	
iei	m**i.ei**	m**iei**	centred
	/m **i e j**/	/m **j e j**/	
	(En. *thousand*, dat.-gen.)	(En. *lamb*, pl.)	
oai	că.s**oa.i**ei	le.**oai**.că	centred
	c @ s **o_X a j** e j/	/l e **o_X a j** c @/	
	(En. *the big house*, dat.-gen.)	(En. *lioness*)	
ioa	că.pr**i.oa**.ră	a.ri.p**ioa**.ră	asc
	/c @ p r **i o_X a** r @/	/a r i p **j o_X a** r @/	
	(En. *doe*)	(En. *little wing*)	
eoa	l**e.oa**i.că	l**eoa**r.că	asc
	/l **e o_X a** j c @/	/l **e_X o_X a** r c @/	
	(En. *lioness*)	(En. *soaking*)	
uea	ta.t**u.ea**.ză	în.șe.**uea**.ză	asc
	/t a t **u e_X a** z @/	/1 n S e **w e_X a** z @/	
	(En. *to tatoo*, pres., 3 sg. and pl.)	(En. *to saddle*, 3 sg. and pl.)	
ioi	v**i.oi**	șo.ri.c**ioi**	asc
	/v **i o j**/	/S o r i c **j o j**/	
	(En. *lively*)	(En *big mouse*)	

were manually corrected in the process, at all levels of lexical information: syllabification, lexical stress marking and phonemic transcription;

- *compound words*: are a problem for RoTTS, which does not deal well with the hyphen in the syllabification step.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

3.2.2. Stress marking correction stage

It is essential, at this level, to review the homographs that are not homophones because, as it can be seen below,[l] stress can distinguish between words implicitly, through lemmas and/or POSes (e.g., 'război') or different morphological variants of the same word (e.g., 'atribui'). It can also influence syllabification and phonemic transcription. Although most of the cases affect two words, there are cases of homography affecting three words: for example, the form 'dudui' can be stressed as: (i) d'udui when it is the second person singular of the present tense of the verb 'a dudui' (En. *to whirr*), (ii) dud'ui when it is the indefinite plural of the noun 'duduie' (En. *madam*), (iii) dudu'i when it is the infinitive or the third person singular past simple form of the same verb 'a dudui' (Băcilă 2011).

The common types of homonymy that introduce ambiguities are as follows:

(1) Lexical homographs
   a. different POSes:
      i. `război (En. war), noun, răzb'oi;`
      ii. `război (En. (to) fight), verb, războ'i`
   b. the same POS, different meanings
      i. `ţarină (En. tsarina), noun, ţar'ină`
      ii. `ţarină (En. cultivated land), noun, ţ'arină`

(2) Lexico-grammatical homographs
   a. same POS
      i. `fotografii (En. photos), fotograf'ii`
      ii. `fotografii (En. photographers), fotogr'afii`
   b. different POSes
      i. `data (En. the date), noun, d'ata`
      ii. `data (En. (to) date) verb, dat'a`

(3) Morphological homographs
   a. different forms in the inflectional paradigm of the same verbal lemma
      i. `atribui (En. (to) assign), verb first or second person singular, present tense, atr'ibui`
      ii. `atribui (En. (to) assign), verb third person singular, past tense infinitive, atribu'i`

3.2.3. Phonemic transcription correction stage

The phonetic alphabet adopted by our dataset is based on the SAMPA notation.[m] The difference between the official SAMPA phonetic notations and our phoneme list lies in our extension of the phonemes inventory as follows:

(1) introducing two notations for transcribing the two possible pronunciations corresponding to letter 'x' (which is not dealt with in SAMPA):/gz/ or/cs/. These notations are in line with the treatment of 'x' as a single consonant in the syllabification phase: for example, the word 'examen' contains the syllables: e.xa.men which correspond to the transcription/e gz a m e n/; if we had transcribed the word as/e g z a m e n/, then the syllabification rule

[l] A comprehensive list of the possible subtypes of the three types of homographs, that is lexical, lexico-morphological and morphological homographs, in this table is made by Băcilă (2011) who considers all parts of speech, as well as all their morphological categories when classifying Romanian homographs.

[m] www.phon.ucl.ac.uk/home/sampa

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

according to which two consonants between two vowel belong to different syllables, that is VCCV~ > ~VC.CV, would not have been observed and an exception should have been formulated;

(2) introducing two new notations to distinguish between the voiceless palatal plosives/k_j/ and the voiced palatal plosive/g_j/, on the one hand, and the voiceless velar plosive/k/ and the voiced velar plosive/g/, on the other hand, as they are different sounds, given their different positions of articulation;

(3) introducing the special notation/je/ for the pronunciation of the letter 'e' when occurring in only two contexts: the initial position in the forms of the personal pronoun and in the forms of the verb 'a fi' (En. *to be*). In all its other occurrences in initial position of a word, *e* should never be pronounced like this.[n]

Table 3 presents the phonemic transcription of the letters and letter groups in Romanian that introduce ambiguities and, therefore, can cause transcription errors. The following rules were derived and implemented for the automatic correction of the phonemic transcription:

(1) **Rules for the letter/sound groups 'ce/ci/ge/gi/che/chi/ghe/ghi'**, concerning the transcription of the final vowels ('e', 'i'):

- **Case I:** the group is a word ending:
  a. *when the group forms a syllable by itself, the final letter ('e'/'i') is a vowel (transcribed/e/ or/i/)*; Examples: tre.**ce** (En. *(to) pass*)/t r e tS e/, ghi.**ci** (En. *(to) guess*)/g_j i tS i/, mer.**ge** (En. *(to) walk, (to) go, (to) function*)/m e r gZ e/, a.mă.**gi** (En. *(to) deceive*)/a m @ gZ i/, u.re.**che** (En. *ear*)/u r e k_j e/, o.**chi** (En. *(to) aim*)/o k_j i/, ve.**ghe** (En. *watch*)/v e g_j e/, zbu.**ghi** (En. *(to) gush*)/z b u g_j i/;
  b. *the group does not form a syllable by itself, the final letter 'i' has 'zero' phonetic value (it is not transcribed)*; Examples: mi**ci** (En. *small*, pl.)/m i tS/, lun**gi** (En. *long*, pl.)/l u n gZ/, o**chi** (En. *eyes*)/o k_j/, un**ghi** (En. *angle*)/u n g_j/, o.bli**ci** (En. *oblique*, pl.)/o b l i tS/;
- **Case II:** the group stands as a syllable ending inside the word:
  a. *the final letter ('e'/'i') is always a vowel (transcribed/e/ or/i/)*; Examples: er.ba.**ce**.e (En. *herbaceous*, fem. sg.)/e r b a tS e e/, sal.**ci**.e (En. *willow*)/s a l tS i e/, **ge**.o.log (En. *geologist*, masc.)/gZ e o l o g/, spon.**gi**.os (En. *spongy*, masc. sg.)/s p o n gZ i o s/, în.**che**.ia (En. *(to) finish*)/1 n k_j e j a/, în.**chi**.na (En. *(to) dedicate, (to) worship*)/1 n k_j i n a/, **ghe**.ţar (En. *glacier*)/g_j e tS a r/, **ghi**.o.cel (En. *snowdrop*)/g_j i o tS e l/;
- **Case III:** the group is inside the syllable;
  a. *the final letter ('e'/'i') is a vowel (transcribed/e/ or/i/), when the group is followed by a consonant*; Examples: **ce**r.ta (En. *(to) scold*)/tS e r t a/, în.**ci**n.ge (En. *(to) heat*)/1 n tS i n gZ e/, **ge**r.man/gZ e r m a n/, ar.**gi**nt (En. *silver*)/a r gZ i n t/, **che**l.tui (En. *(to) spend*, indicative, present)/k_j e l t u j/, **chi**n.gă (En. *strap*)/k_j i n g @/;
  b. *when the group is followed by one or two vowels/semi-vowels, the rules for diphthongs and triphthongs transcription are applied*:
  i. when descendant diphthongs are involved: for example, for the sequence 'cei' in the word 'cercei' (En. *earrings*), we reproduce the/tS/ symbol, followed by the descendant diphthong transcription of 'ei', which is/e j/; in the case of the word 'mijloc'iu' (En. *middle one*), we know that the group of letters 'iu' is a descendant diphthong transcribed as/i w/, because 'i' bears the stress marking, and therefore 'u' is the semi-vowel; Examples: cer.**cei**/tS e r tS e j/, mij.lo.**ci**u/m i Z l o tS i w/, a.po.**ge**u (En. *climax*)/a p o

[n]This special treatment of such cases is meant to show that this pronunciation is the norm, and not a mere allophone of/e/.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

gZ e w/, han.**gi**u (En. *innkeeper*)/h an gZ i w/, în.**che**i (En. *(to) finish*, 1st person, sg.)/1 n k_j e j/, mu.**chi**i (En. *edges*)/m u k_j i j/, pâr.**ghi**i (En. *leverages*)/p 1 r g_j i j/;

ii. for the ascending diphthongs and for ascending and centred triphthongs, the last letter of the group 'ce/ci/ge/gi/che/chi/ghe/ghi' is not transcribed (it has 'zero' value, because in the diphthong or triphthong it is a semi-vowel); Examples: **ce**a.ță (En. *fog*)/tS a ts @/, pi.**ci**or (En. *leg*)/p i tS o r/, **ge**a.nă (En. *eyelid*)/gZ a n @/, **gi**u.va.er (En. *gem*)/gZ u v a e r/, **che**a.mă (En. *(to) call*, imperative, singular)/k_j a m @/, **chi**ar (En. *even*)/k_j a r/, **ghe**a.ță (En. *ice*)/g_j a ts @/, **ghi**oz.dan (En. *shoolbag*)/g_j o z d a n/;

(2) **Rules for diphthong and triphthong transcriptions:** in Romanian, most of these groups can be classified in a deterministic manner, without supplementary context information. The diphthongs can be ascending (semi-vowel + vowel) or descending (vowel + semi-vowel). The triphthongs can be ascending (semi-vowel + semi-vowel + vowel) or centred (vowel + semi-vowel + vowel). The diphthong 'iu' is the only exception: it can be both descending (e.g., in 'fiu'/f i w/, 'hangiu'/h a n gZ i w/, 'mijlociu'/m i Z l o tS i w/) and ascending (e.g., in 'iubit'/j u b i t/, 'iute' (En. *fast*) j u t e/). The ascendance of 'iu' can be identified if 'i' in the diphthong is correctly stressed: if 'i' bares a stress mark, 'iu' is a descending diphthong and otherwise is an ascending one. In Table 4, you can see examples for all the diphthongs and triphthongs in Romanian and also of the same letter group in their hiatus form.

(3) **Rule for the final 'whispered' i:** If the last or the only syllable in the word does not bear stress and it ends with a sequence of the form 'vowel + consonant + (optional consonant) + i', then the final 'i' is transcribed as 'i_0'. Examples: 'primăveri' (En. *springs*)/p r i m @ v e r i_0/, 'beți' (En. *drunk*, pl.)/b e ts i_0/, 'conți' (En. *counts*)/c o n ts i_0/, 'cerbi' (En. *deer*, pl.)/tS e r b i_0/; exceptions from the rule are the groups 'ci/gi/chi/ghi' (for which the rule I.b is applied) and the sequences 'consonant + liquid consonant (/l/ or/r/) +/i/': 'co.dri' (En. *old forests*)/c o d r i/, 'cio.cli' (En. *grave-digger*)/tS o c l i/.

For the words containing the letter or letter groups 'x', 'ki' and 'qu', no automatic correction rules could be determined. Therefore, the entries containing the letter 'x', which can be pronounced as either/ks/ or/gz/, were manually corrected. Manually correcting only the entries for which the word form coincides with the lemma is enough, since the pronunciation of this letter does not change in the inflection process and can be safely extended to its all inflected forms. The words containing the letter groups 'ki' and 'qu' were processed so that the groups be transcribed as/k_j i/ and/k_j/, respectively, to deal with the ambiguities presented in Table 3.

Based on the automatic processes, the derived rules and manual correction, the entire RoLEX dataset was validated and finalised to contain all the linguistic information set forth. Details of its final content are described in the following subsection.

3.3. RoLEX statistics

With the dataset in place, we performed a series of statistics over its lexical components. In the final form, RoLEX contains 330,866 entries and represents the largest phonological validated dataset freely available for Romanian. Table 5 presents an overview of its contents. We add here a remark regarding the number of distinct syllables found in RoLEX. Previous studies of the Romanian syllable distribution (Dinu 2004; Dinu *et al.* 2006), performed on the DOOM dictionary (1982), identified 6496 type syllables. The work in Barbu (2008), also based on DOOM, but coupled with a paradigmatic mechanism of automatic inflectional generation, refers to an extended dataset of 525,530 entries in which 8,600 syllable types were identified. In contrast, our dataset, being a corpus-based one, has particularities that produce 979 new type syllables derived from (i) foreign

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 5. RoLEX dataset statistics

Feature	Count
Number of entries	330,866
Number of distinct lemmas	29,505
Maximum number of forms for a lemma (the verb 'a fi' (En. *to be*))	117
Average number of forms for a lemma	10
Number of content words	328,631
Number of function words	2,235
Number of homographs	58,522
Number of homophones	56,737
Number of distinct phonemes	37
Number of distinct syllables	9,220
Maximum number of syllables per word (21 entries)	11
Average number of letters per syllable	3
Number of letters in the longest word ('electroglotospectrografiilor')	28

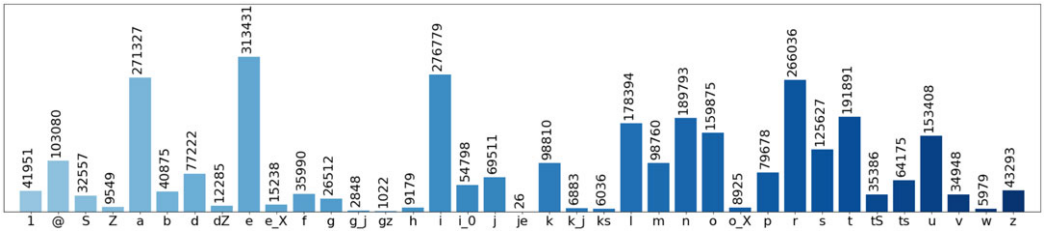

Figure 1. Phoneme counts in RoLEX.

proper nouns that come with specific phonetic properties; (ii) Romanian proper nouns, including some that preserve the old Romanian orthography (namely, the use of the letter 'î' in word internal position); (iii) new and/or borrowed words; (iv) forms occurring in contractions and displaying apheresis; (v) archaisms or regional variants of words; and (vi) interjections.

A detailed statistic of the phoneme counts within RoLEX is shown in Figure 1. The top three most common phonemes being the vowels 'e', 'i' and 'a', followed by the consonants 'r', 't' and 'n'. A separate set of statistics refers to the number of syllables within a word (see Figure 2a) and the position of the stressed syllable within the word (see Figure 2b). The majority of the Romanian words have 3 to 5 syllables, and the most common stress pattern falls on the penultimate or ante-penultimate syllable. Although the lexical stress seems to adhere to a pattern, we will show in the evaluation section that the stress marking poses most problems for the automatic lexical information prediction tool.

In terms of morphological content, Table 6 shows RoLEX's counts for each POS. It can be noticed that the highly flexing POS in Romanian, especially verbs, but also some pronouns, adjectives and nouns, take up 99% of the entire dataset.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

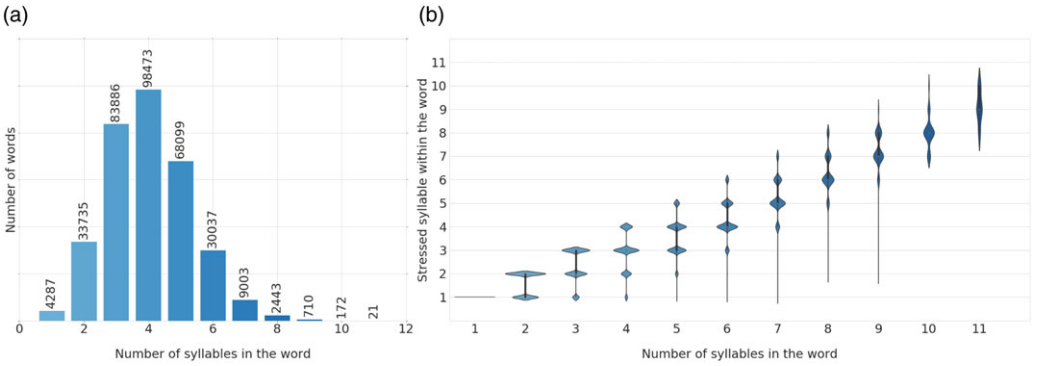

Figure 2. (a) Histogram of the number of syllables per word and (b) violin plot of the stressed syllable position given the number of syllables in a word as computed from the RoLEX dataset.

4. Concurrent lexical information prediction

When such a large high-quality language resource is available, lexical information predictors should be easily and accurately trained. With the recent advancements of core deep learning strategies, as well as deep learning within NLP, predicting a single task at a time can become trivial (for some languages), as well as time and resource consuming. Predicting multiple lexical information at the same time using a single network would be both advantageous and challenging. Such predictors could also exploit the correlations and additional information that would inherently become available in this scenario. As a result, in the rest of the paper we focus on deriving simultaneous lexical information starting from the orthographic form of a context-free word.[o] The selected tasks for the concurrent prediction are as follows: *phonemic transcription*, *syllabification* and *lexical stress marking*. Examples of such input-output pairs are shown in Table 7.

Within this setup, the machine learning algorithm needs to learn a sequence-to-sequence (S2S) mapping. Among the various state-of-the-art neural architectures, CNN (Gehring *et al.* 2017) and attention-based (Vaswani *et al.* 2017) networks have shown the highest accuracy in NLP pipelines (Devlin *et al.* 2019). In the early stages of this study, we first performed a CNN versus Transformer evaluation. However, the CNN results were less accurate than those of the Transformer,[p] so we resumed to using only the latter.

The Transformer architecture is shown in Figure 3 and is composed of an encoder and a decoder structure. Both structures contain a sequence of attention, normalisation and feed forward layers. An important aspect of the Transformer, beneficial to the tasks addressed in this article, is the multi-head attention. By enabling the network to focus on multiple areas of the input sequence, the decoded output can, at each time step, look both into the future and into the past input characters, and adjust the predictions accordingly, yielding a higher accuracy.

The Transformer's hyperparameter selection is based on the strategy introduced in (Stan 2020). The set of hyperparameters which were optimised are shown in Table 8. The optimisation used a randomly selected 150,000 entries subset of RoLEX and evaluated the fitness of the individuals using the word error rate measured for 500 held-out samples. The evolution took place over 10 generations with a population size of 10. This setup does by no mean explore the entire hyperparameter space, yet it allows to evaluate some key topological aspects of the network and prevent overfitting.

[o] We should note here that RoLEX contains only context-free words, and a GOLD standard context corpus with correct annotations is not available for Romanian. Therefore, although context can help disambiguate non-homophone homographs, we could not use contextual information in the prediction at this point.

[p] Previous studies also reported this result on similar tasks (Yolchuyeva *et al.* 2019b).

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 6. RoLEX POS statistics

Tag	POS	Count
V	Verb	129,211
N	Noun	110,232
A	Adjective	89,188
R	Adverb	730
P	Pronoun	345
M	Numeral	340
D	Determiner	268
Y	Abbreviation	251
I	Interjection	156
S	Adposition	75
C	Conjunction	31
T	Article	28
Q	Particle	11

Table 7. Examples of input-output pairs for the concurrent prediction task. Dots mark the syllabification. The lexical stress is marked with an apostrophe before the stressed vowel. The phonemic transcription uses the SAMPA notation

Input	Output	
abandonat	a . b a n . d o . n ’a t	(En. *abandoned*)
bascula	b a s . k u . l ’a	(En. *to swing out*)
ciclopul	tS i . k l ’o . p u l	(En. *the cyclop*)
ăstea	’@ s . t e_X a	(En. *these*)
şchioapa	S k_j o_X ’a . p a	(En. *the limping woman*)

The derived Transformer structure uses 3 encoder units, 4 decoder units, 4 attention heads, a hidden layer of 1024 nodes and an embedding dimension of 128. The embedding weights are randomly initialised before training. The batch size was set to 512, and the Adam optimiser was used to update the weights with an initial learning rate of $2 * 10^{-4}$. After 50 epochs, the learning rate was reduced by a factor of 0.2. An early stopping criterion based on the validation loss over 5 epochs stopped the training process.

5. Evaluation

5.1 Romanian: RoLEX

The evaluation of the newly built RoLEX dataset attempts to answer the following five questions:

[Q1] Is the performance of prediction tools trained on RoLEX better than of those trained on other available Romanian datasets?

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

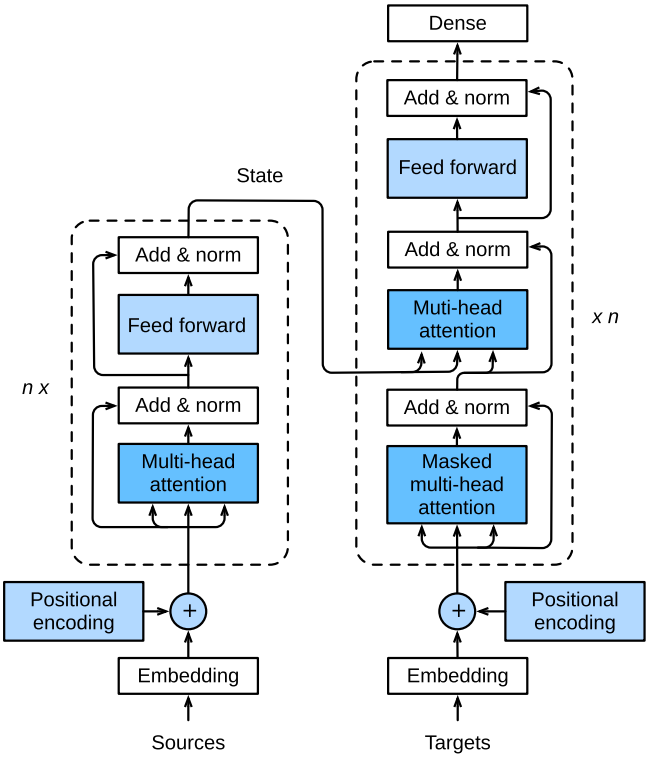

Figure 3. Transformer architecture (Zhang *et al.* 2020).

[Q2] Can the prediction tools trained on a smaller, **randomly selected** subset of entries from RoLEX obtain similar performance measures to those obtained when trained on the entire dataset?
[Q3] Can the prediction tools trained on a smaller, **carefully selected** subset of entries from RoLEX obtain similar performance measures to those obtained when trained on the entire dataset?
[Q4] What is the contribution of each of the three lexical tasks (i.e., phonemic transcription, syllabification and lexical stress assignment) to the global error rates?
[Q5] To what extent do supplemental lexical input features, in the form of POS or morphosyntactic description (MSD) tags, improve the overall accuracy of the prediction tools?

To answer these questions, a 20% randomly selected subset of the RoLEX entries was held out and used in all testing scenarios. `Word error rate` (WER) and `symbol error rate` (SER) were used as objective metrics. The WER was measured as the percentage of incorrectly predicted output sequences. The SER is very similar to the phone error rate, but we would like to make the distinction that the prediction also includes the syllabification and lexical stress symbols. The SER was computed using the Levenshtein distance (Levenshtein 1966) between the predicted and target sequences. For homographs, the pronunciation with the lowest Levenshtein distance was selected. Because the output of the network contains 3 separate types of lexical information, the WER and SER were also computed over the output sequence when removing either the syllable marks, the lexical stress marks, or both. This computation helps us understand which task imposed more learning and prediction problems for the neural network.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 8. Set of genes and gene values used in the evolution strategy. The first column marks the gene ID within the genome

ID	Gene	Values
G1	encoder layers	2, 3, 4
G2	decoder layers	2, 3, 4
G3	embedding dimension	32, 64, 128
G4	attention heads	2, 4
G5	dropout rate	0.01, 0.05, 0.1, 0.15
G6	hidden layer dimension	32, 64, 128, 256, 512, 1024
G7	batch size	32, 64, 128, 256, 512

Table 9. WER and SER measures for individual task predictions

Task	WER	SER
Phonemic transcription	1.60%	0.45%
Syllabification	2.03%	0.48%
Lexical stress marking	4.16%	0.79%

As a preliminary step, the accuracy of the selected neural architecture for each individual task was examined. The results are shown in Table 9 and used the same train-test split as in the following experiments. It can be noticed that the largest error rates are encountered for the lexical stress prediction.

In trying to answer **[Q1]**, we compared the accuracy of the RoLEX-based prediction with the one obtained from the combined information available in MaRePhor, RoSyllabiDict and DEX, while using the same neural architecture. The latter set contains around 72,000 entries and obtained a WER of 10.47%, and a SER of 3.93% for the combined prediction when using only the orthographic form of the word as input. In the same setup, the RoLEX-based prediction halved the error rates of the predictions, with a 5.6% WER, and a 1.97% SER. Given that RoLEX is about five times the size of the MaRePhor-based dataset and more morphologically diverse, the accuracy leap was not unexpected. When also using the MSD information, available in RoLEX and not available in the other resources, the results become highly accurate (3.08% WER and 1.08% SER), with most of the errors pertaining to exceptions. This shows once again the value of extended, manually validated resources, with complex annotation.

A combination of the WER and SER results for answering **[Q2]**, **[Q4]** and **[Q5]** is shown in Figure 4. The results are grouped by the increasing number of randomly selected entries used in the training process. The different colour shades mark the lexical information maintained in the prediction, meaning that the network still predicts the complete lexical information, but we do not take into account all of it. The hatch pattern indicates the information used as input to the neural network: only the orthographic form of the word (Ortho); the orthographic form plus the POS tag (wPOS); or the orthographic form plus the complete MSD tag (wMSD).

For **[Q2]**, it can be noticed that beyond 100,000 entries the accuracy gains seem to plateau, yet there is still a 15% relative WER improvement going from 100,000 entries to the complete dataset in the Ortho:PH + SYLL + ACC setup.

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

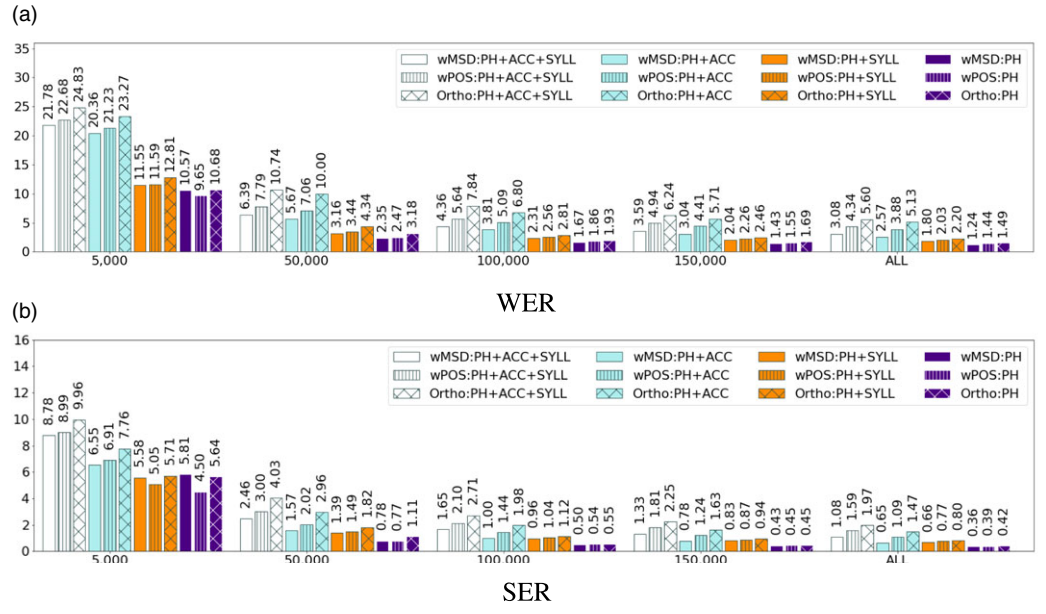

Figure 4. **(a)** WERs and **(b)** SERs for different amounts of randomly selected training samples, evaluated for the complete lexical information prediction (PH + ACC + SYLL); by discarding the syllable information (PH + ACC); by discarding the lexical stress information (PH + SYLL); and by discarding both the syllable and the lexical stress information (PH). Figures also show results of the networks using as input only word forms (Ortho); word forms plus POS tags (wPOS); word forms plus MSD tags (wMSD).

With respect to the influence of each of the three tasks (**[Q4]**), the lexical stress poses the most problems. On average, the lexical stress errors amount to 60% of the overall errors (compare PH + ACC + SYLL with PH + SYLL). This was to be expected for Romanian, as the lexical stress does not adhere to any predefined rules and it is mostly dependent on the word's inflection (DOOM 2005). A similar result was found in (Stan and Giurgiu 2018) and also in the individual task predictions shown in Table 9. Another important aspect to notice in these results for **[Q4]** is the fact that the error rates of the concurrently predicted phonemic transcriptions (Ortho:PH) – when discarding the other tasks – are better than those obtained when the network predicts just this task: 1.60% WER, 0.45% SER (see Table 9). This means that, again as expected, although the network had a more complex learning task, the presence of the other lexical information in the output sequence helps the individual tasks' learning.

In scenario **[Q5]**, the additional lexical information appended to the input in the form of the POS or complete MSD tags should help the network differentiate non-homophone homographs. RoLEX contains only 2000 such type of homographs and we did not envision that the error rate would be significantly impacted by their discrimination. But solving this disambiguation problem can bring more linguistic accuracy to the overall system. Also, when POS and MSD information is used to improve the task's performance, the assumption is that they can compensate for some missing words in the training data. For example, the network can learn to associate certain morphological suffixes (and their specific pronunciations) to certain POSes or MSDs. This assumption holds true across all random dataset partitions and all output sequence tasks presented in Figure 4. MSD systems (wMSD) perform better than the POS systems (wPOS), which in turn are less erroneous than the orthographic ones (Ortho).

Questions **[Q1, Q2, Q4, Q5]** analysed the dataset from the given, available resource perspective. However, the development of such a large resource is extremely time consuming and requires expert linguists to perform the manual annotation. Therefore, in **[Q3]** we investigate if the careful

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

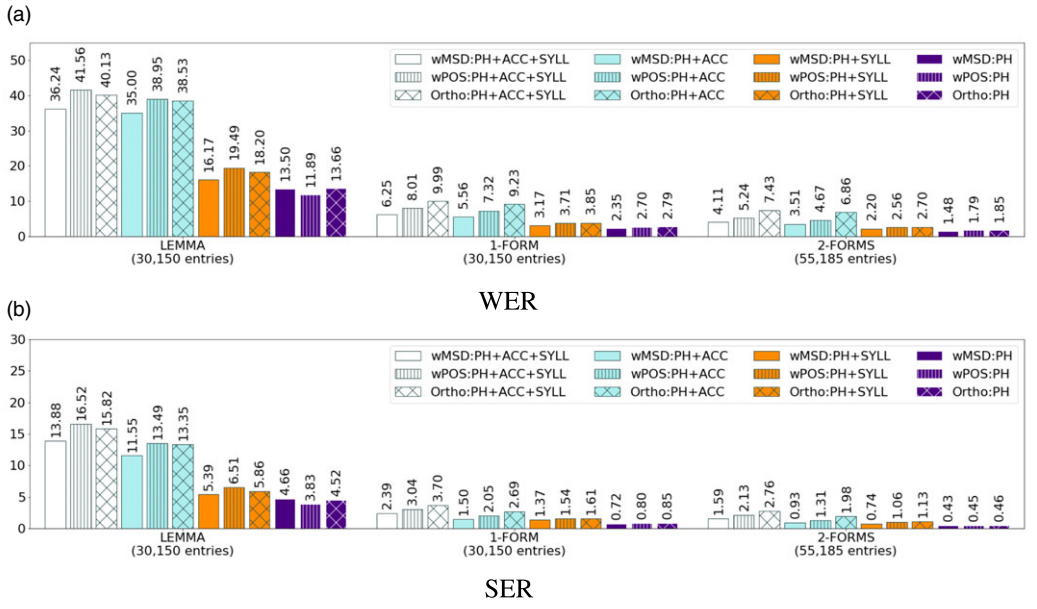

Figure 5 **(a)** WERs and **(b)** SERs at for the `LEMMA`, `1-FORM`, and `2-FORMS` subsets, evaluated for the complete lexical information prediction (`PH + ACC + SYLL`); by discarding the syllable information (`PH + ACC`); by discarding the lexical stress information (`PH + SYLL`); and by discarding both the syllable and the lexical stress information (`PH`). Figures also show results of the networks using as input only word forms (`Ortho`); word forms plus POS tags (`wPOS`); word forms plus MSD tags (`wMSD`).

design and selection of entries can minimise the required manual annotation and validation processes as a more efficient alternative to the validation techniques described in Section 3. Three new subsets of `RoLEX` were generated. The subsets are based on the number and nature of the forms of content words (adjectives, nouns and verbs), which have a rich morphology in Romanian: `LEMMA` subset contains 30,150 entries, corresponding to all forms for the function words and the lemma form for the content words; `1-FORM` subset contains 30,150 entries, corresponding to all forms for the function words and one form for content words, where the selection of the form was performed such that the combined entries ensured the morphological diversity within the corresponding POS category; `2-FORMS` subset contains 55,185 entries and is similar to `1-FORM` but with two forms for each content word entry. The results of the concurrent prediction using these subsets are shown in Figure 5.

The first thing to notice is the very high error rates for the `LEMMA` subset – twice as high as the rates achieved by the randomly selected 5,000 entries (see Scenario [**Q2**]). This can be explained by the very low morphological diversity within the subset. In this case, the POS or MSD tags cannot truly compensate for the lack of morphological variation within the training set. More so, the POS information reduces the accuracy of the prediction. Compare for example the 41.56% WER of `wPOS:PH + ACC + SYLL` vs. 40.13% WER for `Ortho:PH + ACC + SYLL`. One exception is the `wPOS:PH` setup where the POS tags help the phonemic transcription better than the `Ortho` or `wMSD` inputs. However, it seems that the complete MSD tags do aid the concurrent prediction process and lower the WER and SER by approximately 10% relative. We should reiterate the fact that the test set is the same across all evaluation scenarios and includes entries with various morphological forms.

By thoroughly analysing the network's predictions, we discovered that most of the errors are a consequence of a biased learning of lexical stress behaviour. In the `LEMMA` set, more than half of the entries and the majority of the verbs have the lexical stress marking on the last syllable. This feature

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

Table 10. WERs and SERs for the complete lexical information prediction over the augmented CMUDict English dataset

Output sequence	WER	SER
PH + ACC + SYLL	24.11%	4.09%
PH + ACC	23.65%	3.80%
PH + SYLL	22.62%	3.59%
PH	21.98%	3.54%
PH (Yolchuyeva *et al.* 2019a)	22.10%	5.10%

is not characteristic of Romanian's diverse morphological forms. Also, there are numerous errors for syllabification and phonemic transcription in the morphological termination of the words. This means that using only dictionary forms of the entries is not a correct manner to go about selecting the core entries of a lexical dataset.

Looking at the `1-FORM` and `2-FORMS` results in conjunction with the randomly selected subsets, we can see that the careful design of morphologically diverse entries yields performances comparable to those obtained by twice as many random entries: compare the WER of `1-FORM` versus the WER of 50,000 random entries, and the WER of `2-FORMS` to the WER the 100,000 random subset (see Scenario **[Q2]**). These results demonstrate that a strategic morphological selection of the entries substantially reduces the amount of necessary manual validation work for the same target performance. However, the selection process needs to be adapted according to the characteristics of the target language.

5.2. English: CMUDict

The ability to concurrently predict the three lexical tasks in any language using the same neural architecture can enable the development of a flexible multi-lingual framework. We therefore test the Transformer-based structure's feasibility and accuracy for the English CMUDict dictionary, as well. This pronunciation dictionary, developed by the Carnegie Mellon University, consists of more than 135,000 entries, each being associated with its phonemic transcription and lexical stress. The original phonemic and lexical stress transcriptions from CMUDict were combined with the syllabification[q] derived by a method described in (Bartlett, Kondrak, and Cherry 2009). This augmented dataset was used in our experiments and contains 129,420 entries. The results are summarised in Table 10.

The train-validation-test split follows that of Yolchuyeva *et al.* (2019a), with the remark that a fraction of the entries (less than 0.2% of the test set) were not present in the augmented version of the CMUDict. POS/MSD information was not available for the English entries, so that only the concurrent prediction of phonemic transcription, lexical stress and syllabification based on the orthographic representation of the word entries was evaluated. Again, the WERs and SERs of the predicted phonemic transcription (`PH`) when discarding the other lexical information are comparable to the ones obtained by the state-of-the-art methods (Yolchuyeva *et al.* 2019b). As was the case for Romanian, the results show that concurrent task learning can lead to a better performance of the individual tasks – also indicated by van Esch *et al.* (2016).

[q]Available online: https://webdocs.cs.ualberta.ca/~kondrak/cmudict.html

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

6. Conclusions

Creating and testing tools for processing language are to a large extent sustained by the existence of language resources, on which the tools are trained and/or tuned, and against which they are further tested. When a language lacks such a resource (mainly because of the costs involved), alternative, multi-lingual approaches are sought. This article introduced the collection, development, annotation and validation of an extended Romanian lexical dataset, named RoLEX, comprising over 330,000 entries. The dataset is the largest of this kind for Romanian and even the most comprehensive as far as the types of information consistently and systematically encoded are concerned: each entry contains lemma, morphosyntactic information, syllabification, stress and phonetic information.

To test RoLEX's feasibility in deriving automatic lexical annotation tools, we used the dataset to train a concurrent prediction, Transformer-based neural network. The network was set to predict the phonemic transcription, lexical stress and syllabification of a written word (i.e., having its orthographic form as input), or with the additional help of POS tags, or full morphosyntactic descriptions. The evaluation included the analysis of 5 different scenarios which targeted the amount and quality of training data, input augmentation and the cumulative effect of each task in the overall error. The results show very high accuracy for all tasks and are in line with state-of-the-art methods applied to each individual task. We also showed that by carefully selecting data subsets that reflect the morphological diversity of the language, manual validation can be significantly reduced if an incremental setting of validation-training-validation is designed. As future work, we aim to deliver the full prediction system as a freely accessible API, and we already started to use the combined lexical information as input for end-to-end speech synthesis systems.

Acknowledgement. The research presented herein received funding from the Romanian Ministry of Research and Innovation, PCCDI–UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818/73. AS was partially funded by Zevo Technology through project number 21439/28.07.2021.

Competing interests. The authors declare none.

References

Barbu A.-M. (2008). Romanian lexical data bases: Inflected and syllabic forms dictionaries. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech: European Language Resources Association (ELRA), pp. 1937–1941.

Barbu Mititelu V., **Tufiș D. and Irimia E.** (2018). The reference corpus of the contemporary Romanian language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), pp. 1178–1185.

Bartlett S., **Kondrak G. and Cherry C.** (2009). On the syllabification of phonemes. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, Boulder, pp. 308–316.

Băcilă F.-M. (2011). O posibilă clasificare a omografelor românești. *Philologica Banatica* **V**(1), 36–46.

Boroș T., **Dumitrescu S. D. and Pais V.** (2018). Tools and resources for Romanian text-to-speech and speech-to-text applications. *CoRR*, abs/1802.05583.

Chae M., **Park K.**, **Bang J.**, **Suh S.**, **Park J.**, **Kim N. and Park L.** (2018). Convolutional sequence to sequence model with non-sequential greedy decoding for grapheme to phoneme conversion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2486–2490.

Ciobanu A. M., **Dinu A. and Dinu L. P.** (2014). Predicting Romanian stress assignment. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers*, pp. 64–68.

Cucu H., **Buzo A.**, **Besacier L. and Burileanu C.** (2014). SMT-based ASR domain adaptation methods for under-resourced languages: Application to Romanian. *Speech Communication* **56**(8), 195–212.

de Mareüil P. B., **d'Alessandro C.**, **Yvon F.**, **Aubergé V.**, **Vaissière J. and Amelot A.** (2000). A French phonetic lexicon with variants for speech and language processing. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens: European Language Resources Association (ELRA).

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN: Association for Computational Linguistics, pp. 4171–4186.

Diaconescu S.-S., Codirlasu F.-C., Ionescu M., Rizea M.-M., Radulescu M., Minca A. and Fulea S. (2015a). *Fonetica Limbii Romane: Vol. 2 Dictionarul morfologic si fonetic al limbii romane (A-L), Vol. 3 Dictionarul morfologic si fonetic al limbii romane (M-Z)*. Scotts Valley, CA: CreateSpace.

Diaconescu S.-S., Codirlasu F.-C., Ionescu M., Rizea M.-M., Radulescu M., Minca A. and Fulea S. (2015b). *Fonetica Limbii Romane: Vol. 2 Dictionarul morfologic si fonetic al limbii romane (A-L), Vol. 3 Dictionarul morfologic si fonetic al limbii romane (M-Z)*. Scotts Valley, CA: CreateSpace.

Dinu L. (2004). *Despartirea automata in silabe a cuvintelor din limba română. Aplicatii in construcția bazei de date a silabelor limbii române*. Universitatea Bucuresti.

Dinu L., Ciobanu A. M., Chitoran I. and Niculae V. (2014). Using a machine learning model to assess the complexity of stress systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA), pp. 331–336.

Dinu L. and Dinu A. (2006). On the data base of Romanian syllables and some of its quantitative and cryptographic aspects. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa: European Language Resources Association (ELRA), pp. 1795–1798.

Dinu L. P. (2003). An approach to syllables via some extensions of Marcus contextual grammars. *Grammars* **6**(1), 1–12.

Dinu L. P., Niculae V. and Sulea O.-M. (2013). Romanian syllabication using machine learning. In *International Conference on Text, Speech and Dialogue*. Pilsen: Springer, pp. 450–456.

Domokos J., Buza O. and Toderean G. (2012). 100K+ words, machine-readable, pronunciation dictionary for the Romanian language. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. Bucharest: IEEE, pp. 320–324.

DOOM (2005). *The Orthographic, Orthoepic and Morphologic Dictionary of the Romanian Language (DOOM2)*. Bucharest: Univers Enciclopedic.

Dou Q., Bergsma S., Jiampojamarn S. and Kondrak G. (2009). A ranking approach to stress prediction for letter-to-phoneme conversion. In *Proceedings of the Joint Conference of the 47th annual meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, pp. 118–126.

Gehring J., Auli M., Grangier D., Yarats D. and Dauphin Y. N. (2017). Convolutional sequence to sequence learning. In *International Conference on Machine Learning*. Sydney: PMLR, pp. 1243–1252.

Georgescu A.-L., Cucu H. and Burileanu C. (2017). SpeeD's DNN approach to Romanian speech recognition. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. Bucharest: IEEE, pp. 1–8.

Georgescu A.-L., Cucu H., Buzo A. and Burileanu C. (2020). RSC: A Romanian read speech corpus for automatic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, pp. 6606–6612.

Goslin J., Galluzzi C. and Romani C. (2014). PhonItalia: A phonological lexicon for Italian. *Behavior Research Methods* **46**(3), 872–886.

Halpern J. (2022). *Comprehensive Full-Form Lexicon for Arabic NLP and Speech Technology*. Online. Available at https://www.cjk.org/wp-content/uploads/Halpern-LREC2022Paper.pdf 18 July 2022.

Ion R. (2018). TEPROLIN: An extensible, online text preprocessing platform for Romanian. In *Proceedings of the 13th International Conference on Linguistic Resources and Tools for Processing the Romanian Language*, Iași.

Kyparissiadis A., van Heuven W. J., Pitchford N. J. and Ledgeway T. (2017). GreekLex 2: A comprehensive lexical database with part-of-speech, syllabic, phonological, and stress information. *PloS one* **12**(2), e0172493.

Levenshtein V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* **10**, 707.

Lőrincz, B. (2020). Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks. In *24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 176*, pp. 108–117.

Milde B., Schmidt C. and Köhler J. (2017). Multitask sequence-to-sequence models for grapheme-to-phoneme conversion. In *Proceedings of Interspeech 2017*, Stockholm, pp. 2536–2540.

Pearson S., Kuhn R., Fincke S. and Kibre N. (2000). Automatic methods for lexical stress assignment and syllabification. In *Sixth International Conference on Spoken Language Processing*, Beijing.

Peiró-Lilja A. and Farrús M. (2020). Naturalness enhancement with linguistic information in end-to-end TTS using unsupervised parallel encoding. In *Proceedings of Interspeech 2020*, Shanghai, pp. 3994–3998.

Protopapas A., Tzakosta M., Chalamandaris A. and Tsiakoulis P. (2012). IPLR: An online resource for Greek word-level and sublexical information. *Language resources and evaluation* **46**(3), 449–459.

Rao K., Peng F., Sak H. and Beaufays F. (2015). Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, pp. 4225–4229.

Rehm G., Berger M., Elsholz E., Hegele S., Kintzel F., Marheinecke K., Piperidis S., Deligiannis M., Galanis D., Gkirtzou K., Labropoulou P., Bontcheva K., Jones D., Roberts I., Hajič J., Hamrlová J., Kačena L., Choukri K., Arranz V., Vasiļjevs A., Anvari O., Lagzdiņš A., Meļņika J., Backfried G., Dikici E., Janosik M., Prinz K., Prinz C., Stampler S., Thomas-Aniola D., Gómez-Pérez J. M., Garcia Silva A., Berrío C., Germann U., Renals S. and Klejch O. (2020).

https://doi.org/10.1017/S1351324922000419 Published online by Cambridge University Press

European language grid: An overview. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, pp. 3366–3380.

Română A. (1982). *Dicționarul ortografic, ortoepic și morfologic al limbii române*. Bucharest: Editura Academiei Republicii Socialiste România.

Soares A. P., **Iriarte Á.**, **De Almeida J. J.**, **Simões A.**, **Costa A.**, **Machado J.**, **França P.**, **Comesaña M.**, **Rauber A.**, **Rato A.**, et al. (2018). Procura-PALavras (P-PAL): A web-based interface for a new European Portuguese lexical database. *Behavior Research Methods* **50**(4), 1461–1481.

Stan A. (2019). Input encoding for sequence-to-sequence learning of Romanian grapheme-to-phoneme conversion. In *Proceedings of the 10th IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Timisoara.

Stan A. (2020). RECOApy: Data recording, pre-processing and phonetic transcription for end-to-end speech-based applications. In *Proceedings of Interspeech 2020*, Shanghai.

Stan A., **Dinescu F.**, **Țiple C.**, **Meza Ș.**, **Orza B.**, **Chirilă M. and Giurgiu M.** (2017). The SWARA speech corpus: A large parallel Romanian read speech dataset. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. Bucharest: IEEE, pp. 1–6.

Stan A. and Giurgiu M. (2018). A comparison between traditional machine learning approaches and deep neural networks for text processing in Romanian. In *Proceedings of the 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR)*, Iași.

Stan A., **Lőrincz B.**, **Nuțu M. and Giurgiu M.** (2021). The MARA corpus: Expressivity in end-to-end TTS systems using synthesised speech data. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, pp. 85–90.

Stan A., **Yamagishi J.**, **King S. and Aylett M.** (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication* **53**(3), 442–450.

Sun H., **Tan X.**, **Gan J.-W.**, **Liu H.**, **Zhao S.**, **Qin T. and Liu T.-Y.** (2019). Token-level ensemble distillation for grapheme-to-phoneme conversion. In *Proceedings of Interspeech 2019*, Graz, pp. 2115–2119.

Taylor J. and Richmond K. (2020). Enhancing sequence-to-sequence text-to-speech with morphology. In *Proceedings of Interspeech 2020*, Shanghai, pp. 1738–1742.

Toma S.-A. and Munteanu D.-P. (2009). Rule-based automatic phonetic transcription for the Romanian language. In *2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns*. Athens, pp. 682–686.

Toma S.-A., **Stan A.**, **Pura M.-L. and Barsan T.** (2017). MaRePhoR — An open access machine-readable phonetic dictionary for Romanian. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, pp. 1–6.

Toshniwal S. and Livescu K. (2016). Jointly learning to align and convert graphemes to phonemes with neural attention models. In *2016 IEEE Spoken Language Technology Workshop (SLT)*. San Juan: IEEE, pp. 76–82.

Trandabat D., **Irimia E.**, **Barbu Mititielu V.**, **Cristea D. and Tufiș D.** (2012). *The Romanian Language in the Digital Era*. Metanet White Paper Series. Heidelberg: Springer.

van Esch D., **Chua M. and Rao K.** (2016). Predicting pronunciations with syllabification and stress with recurrent neural networks. In *Proceedings of Interspeech 2016*, San Francisco, CA, pp. 2841–2845.

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A. N.**, **Kaiser L.u and Polosukhin I.** (2017). Attention is all you need. In Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S. and Garnett R., (eds), *Advances in Neural Information Processing Systems 30*. Long Beach, CA: Curran Associates, Inc., pp. 5998–6008.

Webster G. (2004). Improving letter-to-pronunciation accuracy with automatic morphologically-based stress prediction. In *Proceedings of Interspeech 2004*, Jeju Island, pp. 2573–2576.

Yao K. and Zweig G. (2015). Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. In *Proceedings of Interspeech 2015*, Dresden, pp. 3330–3334.

Yolchuyeva S., **Németh G. and Gyires-Tóth B.** (2019a). Grapheme-to-phoneme conversion with convolutional neural networks. *Applied Sciences* **9**(6), 1143.

Yolchuyeva S., **Németh G. and Gyires-Tóth B.** (2019b). Transformer based grapheme-to-phoneme conversion. In *Proceedings of Interspeech 2019*, Graz, pp. 2095–2099.

Zeineldeen M., **Zeyer A.**, **Zhou W.**, **Ng T.**, **Schlüter R. and Ney H.** (2020). A systematic comparison of grapheme-based vs. phoneme-based label units for encoder-decoder-attention models. *arXiv preprint, arXiv: 2005.09336*.

Zhang A., **Lipton Z. C.**, **Li M. and Smola A. J.** (2020). *Dive into Deep Learning*. Available at https://d2l.ai

Cite this article: Lőrincz B, Irimia E, Stan A and Barbu Mititelu V. RoLEX: The development of an extended Romanian lexical dataset and its evaluation at predicting concurrent lexical information. *Natural Language Engineering* https://doi.org/10.1017/S1351324922000419