ON THE PRONUNCIATION OF COMMON LEXICA AND PROPER NAMES IN EUROPEAN PORTUGUESE

Isabel M. Trancoso
INESC/IST

M. Céu Viana
CLUL

Fernando M. Silva
INESC/IST

INESC, R. Alves Redol, 9, 1000 Lisboa, Portugal

ABSTRACT

This paper presents some relevant aspects of the pronunciation of proper names and common lexica in European Portuguese. It starts by a brief description of statistical data concerning the occurrence and distribution of graphemes and phonemes for the two corpora and the distinction between different subclasses found in proper names, namely first and last names, toponyms and acronyms. The central theme of this paper is the comparison of the performance of a rule-based method for letter-to-phone conversion for the two corpora with two self-learning methods, one based on a multi-layered neural network and another based on table look-up. The next section is devoted to the study of acronyms and their formation processes and the large variability found in their native pronunciation. Finally, we shall briefly describe some of the problems posed by the nativized pronunciation of foreign names.

1. INTRODUCTION

In general, the performance of letter-to-sound rule systems for proper names is much worse than the one observed for the common lexicon. This fact is not surprizing since the pronunciation of proper names may deviate from the general pronunciation rules and names of foreign origin may show different degrees of adjustment to the language sound structure.

Other sources of difficulty can, however, be found. The orthography of last names can be rather conservative and, as it does not conform anymore to the general orthographic rules, its phonetic interpretation is sometimes misleading. Furthermore, some applications imply the ability of generating correct pronunciations for acronyms which can follow rules significantly different from the ones observed for the common lexicon.

This paper deals with some of these problems for European Portuguese. The following section briefly describes the corpora of proper names and common lexicon we have used, presenting a comparative study of the distribution of graphemes and phones for the two corpora and other relevant statistical data. Next, the problem of letter-to-phone conversion will be addressed. We shall compare the performance of a rule-based method with that of two self-learning procedures, one based on a multi-layered neural network and another based on table look-up. Before concluding, we shall approach the problems presented by the pronunciation of acronyms. A significant part of this work was done in the scope of the LRE European Project Onomastica.

2. CORPORA

The corpora we have used for studying the pronunciation of proper names were developed on the basis of the material provided by the national telecom operator which participates in the Onomastica project (TLP). This material included around 100,000 isolated words belonging to names of persons, streets, places and companies of the cities of Lisbon and Oporto. Four different subcorpora were relevant for this work:

- Names_Phone1 - subcorpus with the most frequent 20000 names;
- Names_Phone2 - subcorpus of $\approx$ 15000 names, extracted from the previous corpus, excluding foreign names, orthographic spelling errors and acronyms;
- Names_Phone3 - subcorpus of $\approx$ 12000 names, extracted from the previous one, excluding company names formed by common words;
- Acro_Phone - subcorpus of $\approx$ 21000 acronyms (a corpus of newspaper text was also used to augment the list of acronyms extracted from the original 100,000 words).

The constitution of all these corpora is similar, as each entry contains a single orthographic form, its frequency of occurrence and its (broad) phonetic transcription. The transcriptions were automatically generated by the letter-to-phone software developed within the joint INESC/CLUL project on text-to-speech synthesis DIXI [3] and later manually processed to correct phonetic values and the locations of stress marks and syllable boundaries. All entries were also manually classified according to its etimlogy and category (first name, surname, street, town or region, common word, company name and acronym).

In order to compare the statistical data derived from these corpora with the corresponding data for the common lexicon, we have also used a similarly structured corpus designated as PF_Phone, developed within the DIXI project. This frequency corpus was based on the PF corpus collected by CLUL [2], build from a set of oral interviews all around the country. It contains about 26000 citation and inflected forms, corresponding to about 750,000 occurrences of words and to more than 3 million graphemes. From the original list of 100,000 words, roughly 50% are unique occurrences. By ordering this list in decreasing order of frequency, the first 13.000 words occur more than 10 times and the first 2700 more than 100 times. Using the first of these subsets, a coverage of about 88% of the full names in the telephone directory of Lisbon is obtained. For Oporto, the coverage is 91% and for the rest of the country about 84%. If the coverage values are computed in terms of isolated names, larger percentages are obtained: 96% (Lisbon and Oporto) and 93% (rest of the country). Hence, the national coverage of the Names_Phone corpora we have used is very significant. Most of the entries in the Names_Phone1 corpus have multiple categories. A significant percentage of surnames can, for instance, be also classified as place name (33%) or first name (16%). Table 1

presents the percentage values computed for each category, successively excluding from the computation the entries which belong to any of the previouly indicated categories. The last entry corresponds to either foreign origin names (4%) or orthographic spelling errors (2%). The PF_Phone

CATEGORY	%
First name	16
Place name	16
Surname	28
Company name (common lexicon)	17
Company name (acronym)	17
Unclassified	6

Table 1: Distribution of entries in Names_Phone1 according to category

and Names_Phone3 have been used for a comparative analysis of the distribution of graphemes and phones in the common lexicon and in proper names, with and without frequency weighting. Proper names show a larger percentage of liquids ("l", mainly), but the differences are not very significant. The distribution of di-graphemes or diphones evidences some small differences in the ordering of the most frequent ones, specially if one takes into account the frequency of functional words or names such as "Maria", for instance. The same can be observed for trigraphemes and triphones. The comparison of the number of different diphones yields a higher value for the common lexicon (813) than for proper names (791). There are 96 diphones in the first corpus which could not be observed in the second one (e.g., "dk", "dv", bS", some diphones typical of verb forms, etc.). And there are 49 diphones in the corpus of proper names which were not observed in the common lexicon. In terms of triphones, similar facts were observed. Without frequency weighting, the distribution of the number of syllables per word in the two corpora is similar. With weighting, however, one can no longer observe for proper names the pattern of decreasing percentages with increasing number of syllables which is typical of the common lexicon of many languages, reflecting the fact that shorter words are easier to pronounce and hence more common. For a detailed study of the distributions of graphemes, phones and syllabic patterns, see [5].

3. LETTER-TO-PHONE CONVERSION

This section compares three types of methods for letter-to-phone conversion: one based on a rule system and two other self-learning methods. Whereas the first type of method is language specific and requires extensive linguistic knowledge, the second one is characterized by the ease with which it can be ported to a new language, the basic requirement being a database of manually transcribed words. Although such a database exists within Onomastica, we decided to train the self-learning methods on the basis of the common lexicon, to make a fairer comparison with the rule-based system. Hence, our comparative tests used two corpora: a subset of the common lexicon (PF_Phone_Test) and the corpus of proper names (Names_Phone2). The remaining subset of the common lexicon, PF_Phone_Train, consisting of 70% randomly chosen entries ($\approx 100,000$ phones), was used for training the self-learning methods. Secondary stress marks were not taken into account in these tests.

3.1. Rule-Based Approach

The rule-based system we have used was previously developed in the scope of the DIXI project. All the code was written in C, directly in the case of the stress assignment rules, and using the SCYLA ("Speech Compiler for Your Language") multi-level rule compiler, developed by CSELT, for the remaining rules. The DIXI system allows different transcription styles and was later modified to include two options for stress mark placement: the original one, before the syllable nuclei and the new one, which conforms to Onomastica standards, before syllable onsets. Syllabification and stress assignment are activated prior to the phonetic transcription module which includes about 200 rules.

The percentage of errors at word level obtained with this rule set is shown in Table 2 for PF_Phone_Test and Names_Phone2, under the heading of RB (Rule-Based). The percentage of correct phonetic transcriptions for the first corpus was 95.5% and the confidence interval is [95.0%,96.0%] with a confidence level of 95%. For the second corpus, the corresponding values are 92.7% [92.3%,93.1%]. The most common type of errors concerns the transcription of graphemes *e* and *o*, to which two different phonological representations can be associated, and the transcription of *x* whose contextual variation is difficult to predict, since it depends, among other facts, on the time the word entered the lexicon. These values allow

CORPUS	PF_Phone_Test			Names_Phone2		
ERROR	RB	NN	TL	RB	NN	TL
Phon. transc.	4.5	7.3	18.7	7.3	12.4	29.5
Prim. stress	0.4	2.7	–	0.4	1.1	–
Syllabification	0.3	0.8	1.0	0.3	1.0	0.8

Table 2: Comparison of rule-based and self-learning procedures. RB - Rule-Based; NN - Neural Net; TL - Table Look-up

us to conclude that in European Portuguese, contrarily to what is referred for some other languages, the letter-to-phone correspondence does not significantly differ for the two types of corpora.

3.2. Neural-Network Approach

Neural networks are characterized by several properties which, because of their analogy with the nervous system, justify their designation: learning capacity, feature extraction, generalization and parallel processing. These properties are obviously involved in the reading process, which is probably one of the reasons why the first papers reporting the application of neural networks to letter-to-phone conversion date back from 1987, when the NETTALK system [4] was presented. As in this pioneering work, the network we have adopted is a conventional multi-layered, feedforward neural network, trained by the backpropagation algorithm.

The learning phase is preceded by an alignment procedure, also based on the software package developed within the DIXI project, which yields about 200 different letter-phone combinations. Phonemic nulls are inserted to account for graphemes with no phonetic realization (the initial "h" in Portuguese, for instance). The definition of graphemic nulls is also possible although, in our case, it was avoided by the alternative definition of new symbols for sequences of phones corresponding to single graphemes. Each network input pattern is based on one grapheme and its context provided by nearby graphemes. The desired network output is the phone aligned with the input character. Several network architectures and context lengths have been tested. The one which yielded best results so far is illustrated in Fig.1. The input layer consists

of 11 clusters of neurons, one cluster for each grapheme: the one to be transcribed, 3 graphemes to its left and 7 graphemes to its right, from which only 5 are used for phonetic transcription, the two last ones being exclusively used for stress assignment. Each grapheme is encoded by a group of 36 neurons, to account for all the 35 different graphemes with corresponding diacritics and also the word boundary mark. Hence, the total number of input neurons is 396. The hidden layer is structured into 5 clusters of 3 graphemes and 2 clusters of 2 graphemes each. The latter include the grapheme to be transcribed and either the one immediately to it left or right. Each cluster consists of 20 neurons, which amounts to a total of 140 hidden units. There are 47 output neurons, one for each of the 45 different phones (including the phonemic null and the compound phones), and two to encode the primary stress mark and the syllable boundary. Shared weights have been adopted in order to reduce the number of weights to be adjusted (11087 weights for 21167 synapses). There is also one direct connection betwen input and output. After 8 epochs, the error percentage at segment level was already down to 1.5%, reaching 1% at the end of 40.

The results obtained at word level for two test corpora are shown in Table 2, under the heading of NN (Neural Net). Multiple stress marks were attributed to many words, a problem which was solved by simple post-processing. It is interesting to notice that a significant percentage of the words in which the rule-based method fails was also wrongly transcribed by the neural network (59% of the cases for the first test corpus, and 74% for the second one). Moreover, roughly half of these wrongly transcribed words are transcribed in the same way by both approaches (44% and 56%, respectively, for the two corpora). In fact, the most frequent errors of the neural network are also concerned with the *e*, *o* and *x* graphemes, although not necessarily in the same words. The network, however, shows some difficulties in dealing with vowel nasalization: nasal consonants in syllable final positions do not always nasalize the preceding vowel, as they should, and although not so often, they can erroneously nasalize the preceding vowel when associated with the onset of the next syllable. Vowel nasalization can also be wrongly triggered by /l/ in syllable final positions. Other problem concerns vowel raising which, in European Portuguese, can only occur in unstressed syllables. The network, however, sometimes raises vowels in stressed syllables and fails to raise some unstressed ones. It also has difficulties in dealing with diphthongization. Some of the generalization problems are due to the small representativity of the corresponding grapheme sequences in the training corpus.

3.3. Table Look-Up Approach

The results presented in this section were obtained using the self-learning table look-up software package SELEGRAPH developed by the Danish partner of the Onomastica project (IES) [1]. The main difference between this type of self-learning approach and the neural network approach described before is the lack of generalization capabilities in table look-up approaches, a disadvantage that is to some extent counter-weighted by their much faster training procedure. Table look-up approaches are trained on the basis of paired grapheme-phone strings, dynamically determining which left and right contexts are minimally sufficient to be able to map any of the graphemes to the correct phone with absolute certainty. The table look-up training is preceded by two phases: the alignment phase, just as for the neural network approach, and the computation of mutual information, to determine how many context graphemes to include and the ordering in which they should be considered. The training results in tree-structured statistics, for each grapheme in a given context, of the number of occurrences of each possible phone. Default mappings are used for ambiguous conversions and unseen words with grapheme sequences not present in the training corpus.

The results for the two test corpora are shown in Table 2, under the heading TL. The percentage of primary stress mark assignment errors is not shown for TL, as many words were assigned either no stress mark (24% and 28%, respectively for the PF and Names corpora), or two of them (13% and 12%). For this approach, no post-processing scheme was implemented yet. An analysis of the most typical phonetic transcription errors evidences the lack of generalization capability and shows all the error types described for the neural network, although much more frequently. Our rule-based system avoids vowel raising errors by performing stress assignment before phonetic transcription. Likewise, the problems of vowel nasalization and diphthongization are avoided by placing syllable marks prior to phonetic transcription. Most of the errors made by the self-learning methods, therefore, could be avoided if a similar approach was adopted either by simple pre-processing or separate subnetworks.

4. ACRONYMS

Acronyms constitute 38% of the most frequent 50,000 entries of the original TLP database. They considerably differ from the common lexicon in two significant ways. The first one is the significant drop in the performance of both rule-based and self-learning methods for this category: 43% phonetic transcription errors for the rule-based method and 51% for the neural network, with 41% common errors. The second is the fact that their pronunciation by native speakers also shows considerable variation. This was illustrated by a reading test using 10 speakers and 100 randomly selected entries of Acro_Phone, almost totally unknown from the speakers, which only showed 37% of pronunciation agreement.

In the building up of acronyms, common affixes, words, roots, first names, last names, toponyms and almost any possible truncation of those are used, combined with each other, with foreign words, and with word endings specific to this category. With a list of only 660 elements, a significant coverage of the Acro_Phone corpus was reached: 15% of the entries are totally covered by the concatenation of elements (2% with overlap); 50% of the entries are partially covered (23% show the element in the initial position, 20% in the final position and 7% in both positions; the remaining 35% are not covered by this short list.

Two different kinds of formation processes are used: word formation (including compounding and derivation), and word creation (including acronymy in the strict sense, blending and abbreviation proper). Most of the inconsistency found in the pronunciation of acronyms appears to be related to compounding and concerns vowel raising and the pronunciation of *s* and *r*. In the common lexicon, two kinds of compounds can be found [6] that are treated differently in terms of orthography and pronunciation: word compounds may have as many non-raised vowels as their constituents; root-compounds also have, besides those, a binding vowel, "o" or "i", which, in the first case, is not raised either. Graphically, these two types are distinguishable by the fact that the first ones are written as separate words (often with hyphens), and the second ones are writ-

ten as a single word.
Compounding is a very frequent process in the building up of acronyms, occurring more often than in the lexicon. In this category, however, both types of compounds are always coined as single words which is a source of ambiguity in their pronunciation. One of the most striking cases concerns the pronunciation of the binding vowel "o" which is identical to the male gender mark (e.g.: *globomar* can be pronounced as a root compound [ˋglo.bɔ´mar] or word compound [ˋglo.bu´mar]). The same type of ambiguity, however, arises when the first constituent ends in "a". In this case, the acronym should be analyzed as a word compound, but is often interpreted as a simple word. This explains the inconsistency in the pronunciation of "s" and "r" in the beginning of a non-initial constituent (e.g.: *alfasom* can be pronounced as [ˋaɬ.fɐ´sõ] or [aɬ.fɐ´zõ] - *alfa* + *som*). These facts, together with the variation found in the reading test for the first and second readings of the same speaker, suggests that the recognition of words or roots within words is not part of the task of reading Portuguese. This means that roots may not be treated as such, but rather as "affixes". When creation processes are concerned, a considerable amount of variation may also be found. If an acronym is not recognized as such, what often occurs in first readings before the task is identified, normal pronunciation rules apply. Otherwise, speakers may try to associate meaning to truncated forms and read them as compounds or, most probably, they do not raise any vowel in pre-stressed position. This last rule is also applied to derived forms with endings specific to company names (e.g. "ex", "ux"). There is no clear distinction among the different creation processes, all of them being able to lead to the same graphemic strings. However, while acronyms in a strict sense and blends are created to be read, *siglae* (abbreviation proper), can lead to strings which are either read or spelt. In order to predict their pronunciation, a specific set of rules is being designed based on syllabic weighting together with preferred syllabic sequences.
Automatic transcription methods do not take into account the variation problems we have described. Their performance is also affected by other causes, such as the frequent omission of diacritics, and the use of foreign constituents not easily recognized as such. Stress assignment errors are also quite common, as a considerable number of acronyms ends in strings that are not present in the common lexicon (e.g., "ax", "trans").

5. FOREIGN NAMES

The pronunciation of foreign names has been treated differently for the names appearing in our national directory and for the ones provided by other partners. For the latter, in fact, the knowledge of the names' etymology allows some systematic processing. Two transcriptions were provided for each name, corresponding to two different degrees of the knowledge of the foreign language. The first one corresponds to an almost total ignorance of the language and the second one to an average knowledge. The difference between the two is quite significant for romance languages, English and German, but very small for the other languages.
Both transcriptions are first generated automatically by simple scripts (one for each different language) and then manually corrected. The automatic generation is done by substituting isolated graphemes (non-Portuguese characters, for instance) and grapheme sequences. For some languages (as the romance ones, for instance), these scripts are activated prior to our grapheme-to-phoneme transcription rules. For other languages, the scripts may directly produce transcriptions.
Contrarily to the touristic names provided by other partners, the foreign names appearing in our national directory have been processed as all the other Portuguese entries. That is, they have been first pre-processed by a script which slightly modifies the original orthography, then automatically processed through our transcription system, whose output is then manually corrected by one or more transcribers. The script is common to all the etymologies and includes some category-sensitive rules which modify the original orthography in order to to take into account some frequent problems in both acronyms and foreign names. The first set of rules deals with double consonants and the second with ending consonants.
Double consonants are very frequent in foreign names, but not so much in Portuguese. In our language, in fact, the only pairs of identical consecutive consonants are *rr* and *ss*, although old orthographies still maintain other double consonants. Very frequently, people know how to pronounce *ll* in some romance languages, but the common trend for the other double consonants is to pronounce them as a single one. Hence, the first set of rules implements this modification.
The second set of pre-processing rules has been developed for dealing with the problem of ending consonants in polyssylabic words. In European Portuguese, most words end in a vowel, *s* (plural mark), *m* (plural verb form) or *l*, *r* and *z*. However, foreign names and acronyms frequently end in *b,c,d,f,g,j,k,p,q,t,v*. In these cases, there is a marked tendency towards pronouncing these sounds as if an *e* had been added in a final position, thus forming a new ending syllable with schwa [ə] and causing the syllable before that one to be stressed, according to our normal stress rule (e.g. *unicef* will be pronounced as [u.ni´sɛ.fə] instead as [u´ni.səf] as dictated by the rule system). The first set of orthography modifying rules thus adds an *e* to unusual final consonants.
For many foreign names, two pronunciations are indicated. Sometimes, this means that the transcriber ignores the origin of the name and provides two transcriptions, one for each origin. This frequently happens with some names including the grapheme *w*, for instance, which does not occur in Portuguese, and where one transcription is provided assuming an English origin and another one assuming a German one (e.g. *kiwi* [kiw´i] or [ki´vi]). More often, however, the two transcriptions mean two alternative extreme pronunciations among which a range of possible pronunciations may be found. In fact, the most troublesome problem with transcribing foreign names is the large variability in pronunciation that can be found. Moreover, the problem becomes more acute as the reader is faced with less familiar grapheme sequences which makes the transcription of romance languages like Spanish, Italian and French easier and less variable than more 'distant' languages.
Vowels constitute one of the largest sources of variability. The greatest problems are posed by the transcriptions of *e* and *o*, when one assumes no knowledge of the foreign language. In fact, these graphemes may have multiple pronunciations in Portuguese that cannot always be predicted by rule, thus providing a native speaker with multiple choices. Vowel sequences uncommon in Portuguese are also troublesome (*aa, ea, ee, ii, oa, oe, oo*, etc.). With consonants, the variability problem when one assumes no knowledge of the foreign language is not so critical, although one faces multiple choices with some graphemes (*h,s,g,r*, in particular). It is also interesting to notice that,

although there are many consonant clusters which are impossible in Portuguese (*sv*, for instance), their pronunciation is very typical, due to vocalic deletion (as in *severo*, for the above example).
A significant cause of variability is the pronunciation of *m* and *n*. In Portuguese, when these two consonants appear in the syllable coda, they nasalize the preceding vowel. The orthography rule, however, imposes an *m* to be used before *p* or *b* and an *n* before the remaining consonants. This rule is very frequently broken in foreign names, where, moreover, *n* often appears in word endings, which is rare in Portuguese. These facts, combined with the common knowledge of the non-existence of the corresponding nasalized vowels in some foreign languages, may cause the nasal consonant to be pronounced instead of triggering nasalization and it is difficult to predict the most common pronunciation.
Besides the variability in phonetic transcription, there is also a great deal of variability in terms of stress assignment. This is specially true with foreign names ending in *i*, and *r*. In Portuguese, the first ending is only frequent in verb forms and causes the last syllable to be stressed. This ending, however, is very frequent in Italian names and, when it is recognized as such, the corresponding syllable is not stressed. When the origin is not recognized as Italian, the location of the stressed syllable varies very much. The same happens when the name ends in *r*, which, in Portuguese, also indicates a verb form, and causes the last syllable to be stressed. In foreign origin names, this ending syllable is not frequently stressed by Portuguese speakers. This fact, combined with the above mentioned variability in the pronunciation of vowels makes it difficult to predict the pronunciation of names ending in *er* and *or* specially by a Portuguese speaker.
The problem of placing syllabic boundaries in foreign names is the most difficult one. As mentioned before, it appears that the recognition of words or roots within words is not usually part of the task of reading Portuguese. This makes it difficult to place syllabic boundary marks in foreign compound names. The presence of large consonantal clusters in the transcriptions is also troublesome in terms of syllabic division, since these clusters occur only in Portuguese when vowel deletion is applied and in these circumstances, syllabic boundary placement is still an open area for research in our language.

6. CONCLUSIONS

European Portuguese does not show significant pronunciation differences between the common lexicon and proper names, if acronyms are not taken into account. Self-learning letter-to-phone conversion methods based on neural networks have shown the potential to perform as well as rule-based systems, provided that large aligned databases are available. The table look-up approach did not yield so good results. However, many of the current transcription errors of both self-learning approaches may probably be avoided with separate processing of stress and syllable marks. Many aspects of the performance of neural networks remained to be explored, namely the analysis of the activation patterns to determine functional groupings whose comparison with the ones predicted by linguistic models may raise very interesting questions.
One of the main challenges of the Onomastica project as far as the Portuguese language is concerned is the pronunciation of acronyms for which the current automatic transcription systems do not yield good results. This fact, together with the large variability found in their pronunciation justifies a more in-depth study of their formation processes. Another interesting challenge is the pronunciation of foreign names. Our experience in the Onomastica project also reveals a large amount of variability in nativized pronunciations and more work is needed to characterize its main sources.

REFERENCESS

References

[1] O. Andersen and P. Dalsgaard, "A Self-Learning Approach to Transcription of Danish Proper Names", Proc. ICSLP'94, Yokohama, Japan, Sept. 94.

[2] F. Nascimento, L. Marques and L. Segura, "Português Fundamental: Métodos e Documentos", INIC-CLUL, Lisbon, 1987.

[3] L. Oliveira, C. Viana and I. Trancoso, "A rule based text-to-speech system for Portuguese", Proc. Int. Conf. on Acoustics, Speech and Signal Processing, San Francisco, vol. II, pp. 73-76, March 1992.

[4] T. Sejnowski and T. Rosenberg, "Parallel networks that learn to pronounce English text", Complex Systems, pp. 145-168, 1987.

[5] C. Viana et. al., "Sobre a pronúncia de nomes próprios, siglas e acrónimos em Português Europeu", Congresso Internacional sobre o Português, Lisbon, April 1994.

[6] A. Villalva, "Compounding in Portuguese", Rivista di Linguistica, Vol. 4, no. 1, pp. 201-219, 1992.

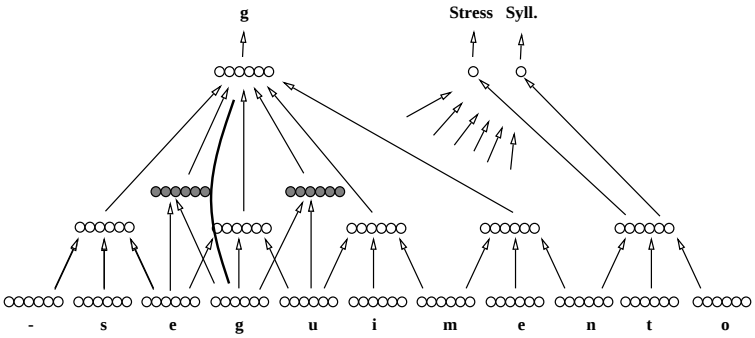

Figure 1: Architecture of the multi-layered neural network