How spoken language corpora can refine current speech motor training methodologies

Daniil Umanski, Niels O. Schiller
Leiden Institute for Brain and Cognition
Leiden University, The Netherlands
daniil.umanski@gmail.com
N.O.Schiller@hum.leidenuniv.nl

Federico Sangati
Institute for Logic,
Language and Computation
University of Amsterdam, the Netherlands
f.sangati@uva.nl

Abstract

The growing availability of spoken language corpora presents new opportunities for enriching the methodologies of speech and language therapy. In this paper, we present a novel approach for constructing speech motor exercises, based on linguistic knowledge extracted from spoken language corpora. In our study with the Dutch Spoken Corpus, syllabic inventories were obtained by means of automatic syllabification of the spoken language data. Our experimental syllabification method exhibited a reliable performance, and allowed for the acquisition of syllabic tokens from the corpus. Consequently, the syllabic tokens were integrated in a tool for clinicians, a result which holds the potential of contributing to the current state of speech motor training methodologies.

1 Introduction

Spoken language corpora are often accessed by linguists, who need to manipulate specifically defined speech stimuli in their experiments. However, this valuable resource of linguistic information has not yet been systematically applied for the benefit of speech therapy methodologies. This is not surprising, considering the fact that spoken language corpora have only appeared relatively recently, and are still not easily accessible outside the NLP community. Existing applications for selecting linguistic stimuli, although undoubtedly useful, are not based on spoken language data, and are generally not designed for utilization by speech therapists per se (Aichert et al., 2005). As a first attempt to bridge this gap, a mechanism is proposed for utilizing the relevant linguistic information to the service of clinicians. In coordination with speech pathologists, the domain of speech motor training was identified as an appropriate area of application. The traditional speech motor programs are based on a rather static inventory of speech items, and clinicians do not have access to a modular way of selecting speech targets for training.

Therefore, in this project, we deal with developing an interactive interface to assist speech therapists with constructing individualized speech motor practice programs for their patients. The principal innovation of the proposed system in regard to existing stimuli selection applications is twofold: first, the syllabic inventories are derived from spoken word forms, and second, the selection interface is integrated within a broader platform for conducting speech motor practice.

2 Principles of speech motor practice

2.1 Speech Motor Disorders

Speech motor disorders (SMD) arise from neurological impairments in the motor systems involved in speech production. SMD include acquired and developmental forms of dysarthria and apraxia of speech. Dysarthria refers to the group of disorders associated with weakness, slowness and inability to coordinate the muscles used to produce speech (Duffy, 2005). Apraxia of speech (AOS) is referred to the impaired planning and programming of speech (Ziegler , 2008). Fluency disorders, namely stuttering and cluttering, although not always classified as SMD, have been extensively studied from the speech motor skill perspective (Van Lieshout et al., 2001).

2.2 Speech Motor Training

The goal of speech therapy with SMD patients is establishing and maintaining correct speech motor routines by means of practice. The process of learning and maintaining productive speech motor skills is referred to as speech motor training.

Proceedings of the ACL 2010 Student Research Workshop, pages 37–42,
Uppsala, Sweden, 13 July 2010. ©2010 Association for Computational Linguistics

An insightful design of speech motor training exercises is crucial in order to achieve an optimal learning process, in terms of efficiency, retention, and transfer levels (Namasivayam, 2008).

Maas et al. (2008) make the attempt to relate findings from research on non-speech motor learning principles to the case of speech motor training. They outline a number of critical factors in the design of speech motor exercises. These factors include the training program structure, selection of speech items, and the nature of the provided feedback.

It is now generally agreed that speech motor exercises should involve simplified speech tasks. The use of non-sense syllable combinations is a generally accepted method for minimizing the effects of higher-order linguistic processing levels, with the idea of tapping as directly as possible to the motor component of speech production (Smits-Bandstra et al., 2006) .

2.3 Selection of speech items

The main considerations in selecting speech items for a specific patient are functional relevance and motor complexity. Functional relevance refers to the specific motor, articulatory or phonetic deficits, and consequently to the treatment goals of the patient. For example, producing correct stress patterns might be a special difficulty for one patient, while producing consonant clusters might be challenging for another. Relative motor complexity of speech segments is much less defined in linguistic terms than, for example, syntactic complexity (Kleinow et al., 2000). Although the part-whole relationship, which works well for syntactic constructions, can be applied to syllabic structures as well (e.g., 'flake' and 'lake'), it may not be the most suitable strategy.

However, in an original recent work, Ziegler presented a non-linear probabilistic model of the phonetic code, which involves units from a sub-segmental level up to the level of metrical feet (Ziegler , 2009). The model is verified on the basis of accuracy data from a large sample of apraxic speakers, and thus provides a quantitive index of a speech segment's motor complexity.

Taken together, it is evident that the task of selecting sets of speech items for an individualized, optimal learning process is far from obvious, and much can be done to assist the clinicians with going through this step.

3 The role of the syllable

The syllable is the primary speech unit used in studies on speech motor control (Namasivayam, 2008). It is also the basic unit used for constructing speech items in current methodologies of speech motor training (Kent, 2000). Since the choice of syllabic tokens is assumed to affect speech motor learning, it would be beneficial to have access to the syllabic inventory of the spoken language. Besides the inventory of spoken syllables, we are interested in the distribution of syllables across the language.

3.1 Syllable frequency effects

The observation that syllables exhibit an exponential distribution in English, Dutch and German has led researchers to infer the existence of a 'mental syllabary' component in the speech production model (Schiller et al., 1996). Since this hypothesis assumes that production of high frequency syllables relies on highly automated motor gestures, it bears direct consequences on the utility of speech motor exercises. In other words, manipulating syllable sets in terms of their relative frequency is expected to have an effect on the learning process of new motor gestures. This argument is supported by a number of empirical findings. In a recent study, Staiger et al. report that syllable frequency and syllable structure play a decisive role with respect to articulatory accuracy in the spontaneous speech production of patients with AOS (Staiger et al., 2008). Similarly, (Laganaro, 2008) confirms a significant effect of syllable frequency on production accuracy in experiments with speakers with AOS and speakers with conduction aphasia.

3.2 Implications on motor learning

In that view, practicing with high-frequency syllables could promote a faster transfer of skills to everyday language, as the most 'required' motor gestures are being strengthened. On the other hand, practicing with low-frequency syllables could potentially promote plasticity (or 'stretching') of the speech motor system, as the learner is required to assemble motor plans from scratch, similar to the process of learning to pronounce words in a foreign language. In the next section, we describe our study with the Spoken Dutch Corpus, and illustrate the performed data extraction strategies.

4 A study with the Spoken Dutch Corpus

The Corpus Gesproken Nederlands (CGN) is a large corpus of spoken Dutch[1]. The CGN contains manually verified phonetic transcriptions of 53,583 spoken forms, sampled from a wide variety of communication situations. A spoken form reports the phoneme sequence as it was actually uttered by the speaker as opposed to the canonical form, which represents how the same word would be uttered in principle.

4.1 Motivation for accessing spoken forms

In contrast to written language corpora, such as CELEX (Baayenet al., 1996), or even a corpus like TIMIT (Zue et al., 1996), in which speakers read prepared written material, spontaneous speech corpora offer an access to an informal, unscripted speech on a variety of topics, including speakers from a range of regional dialects, age and educational backgrounds.

Spoken language is a dynamic, adaptive, and generative process. Speakers most often deviate from the canonical pronunciation, producing segment reductions, deletions, insertions and assimilations in spontaneous speech (Mitterer, 2008). The work of Greenberg provides an in-depth account on the pronunciation variation in spoken English. A detailed phonetic transcription of the Switchboard corpus revealed that the spectral properties of many phonetic elements deviate significantly from their canonical form (Greenberg, 1999).

In the light of the apparent discrepancy between the canonical forms and the actual spoken language, it becomes apparent that deriving syllabic inventories from spoken word forms will approximate the reality of spontaneous speech production better than relying on canonical representations. Consequently, it can be argued that clinical applications will benefit from incorporating speech items which optimally converge with the 'live' realization of speech.

4.2 Syllabification of spoken forms

The syllabification information available in the CGN applies only to the canonical forms of words, and no syllabification of spoken word forms exists. The methods of automatic syllabification have been applied and tested exclusively on canonical word forms (Bartlett, 2007). In order to obtain the syllabic inventory of spoken language per se, a preliminary study on automatic syllabification of spoken word forms has been carried out. Two methods for dealing with the syllabification task were proposed, the first based on an n-gram model defined over sequences of phonemes, and the second based on statistics over syllable units. Both algorithms accept as input a list of possible segmentations of a given phonetic sequence, and return the one which maximizes the score of the specific function they implement. The list of possible segmentations is obtained by exhaustively generating all possible divisions of the sequence, satisfying the condition of keeping exactly one vowel per segment.

4.3 Syllabification Methods

The first method is a reimplementation of the work of (Schmid et al., 2007). The authors describe the syllabification task as a tagging problem, in which each phonetic symbol of a word is tagged as either a syllable boundary ('B') or as a non-syllable boundary ('N'). Given a set of possible segmentations of a given word, the aim is to select the one, viz. the tag sequence $\hat{b}_1^n$, which is more probable for the given phoneme sequence p_1^n, as shown in equation (1). This probability in equations (3) is reduced to the joint probability of the two sequences: the denominator of equation (2) is in fact constant for the given list of possible syllabifications, since they all share the same sequence of phonemes. Equation (4) is obtained by introducing a Markovian assumption of order 3 in the way the phonemes and tags are jointly generated

$$\begin{aligned}
\hat{b}_1^n &= \arg\max_{b_1^n} P(b_1^n|p_1^n) && (1)\\
&= \arg\max_{b_1^n} P(b_1^n, p_1^n)/P(p_1^n) && (2)\\
&= \arg\max_{b_1^n} P(b_1^n, p_1^n) && (3)\\
&= \arg\max_{b_1^n} \prod_{i=1}^{n+1} P(b_i, p_i|b_{i-3}^{i-1}, p_{i-3}^{i-1}) && (4)
\end{aligned}$$

The second syllabification method relies on statistics over the set of syllables unit and bigram (bisegments) present in the training corpus. Broadly speaking, given a set of possible segmentations of a given phoneme sequence, the algorithm, selects the one which maximizes the presence and frequency of its segments.

[1](see http://lands.let.kun.nl/cgn/)

Corpus	Phonemes		Syllables	
	Boundaries	Words	Boundaries	Words
CGN_Dutch	98.62	97.15	97.58	94.99
CELEX_Dutch	99.12	97.76	99.09	97.70
CELEX_German	99.77	99.41	99.51	98.73
CELEX_English	98.86	97.96	96.37	93.50

Table 1: Summary of syllabification results on canonical word forms.

4.4 Results

The first step involved the evaluation of the two algorithms on syllabification of canonical word forms. Four corpora comprising three different languages (English, German, and Dutch) were evaluated: the CELEX2 corpora (Baayenet al., 1996) for the three languages, and the Spoken Dutch Corpus (CGN). All the resources included manually verified syllabification transcriptions. A 10-fold cross validation on each of the corpora was performed to evaluate the accuracy of our methods. The evaluation is presented in terms of percentage of correct syllable boundaries[2], and percentage of correctly syllabified words.

Table 1 summarizes the obtained results. For the CELEX corpora, both methods produce almost equally high scores, which are comparable to the state of the art results reported in (Bartlett, 2007). For the Spoken Dutch Corpus, both methods demonstrate quite high scores, with the phoneme-level method showing an advantage, especially with respect to correctly syllabified words.

4.5 Data extraction

The process of evaluating syllabification of spoken word forms is compromised by the fact that there exists no gold annotation for the pronunciation data in the corpus. Therefore, the next step involved applying both methods on the data set and comparing the two solutions. The results revealed that the two algorithms agree on 94.29% of syllable boundaries and on 90.22% of whole word syllabification. Based on the high scores reported for lexical word forms syllabification, an agreement between both methods most probably implies a correct solution. The 'disagreement' set can be assumed to represent the class of ambiguous cases, which are the most problematic for automatic syllabification. As an example, consider the following pair of possible syllabification, on which the two methods disagree: 'bEl-kOm-pjut' vs 'bEl-kOmp-jut'[3].

Motivated by the high agreement score, we have applied the phoneme-based method on the spoken word forms in the CGN, and compiled a syllabic inventory. In total, 832,236 syllable tokens were encountered in the corpus, of them 11,054 unique syllables were extracted and listed. The frequencies distribution of the extracted syllabary, as can be seen in Figure 1, exhibits an exponential curve, a result consistent with earlier findings reported in (Schiller et al., 1996). According to our statistics, 4% of unique syllable tokens account for 80% of all extracted tokens, and 10% of unique syllables account for 90% respectively. For each extracted syllable, we have recorded its structure, frequency rank, and the articulatory characteristics of its consonants. Next, we describe the speech items selection tool for clinicians.

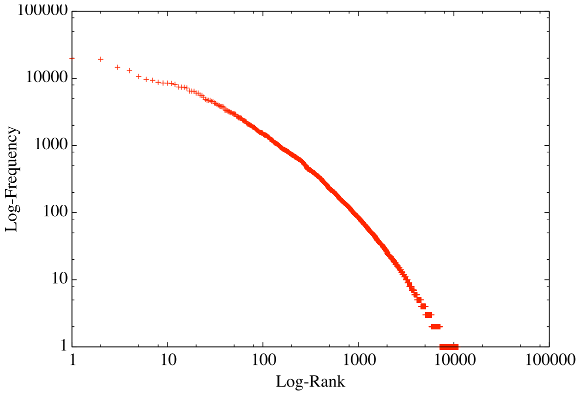


Figure 1: Syllable frequency distribution over the spoken forms in the Dutch Spoken Corpus.
The x-axis represents 625 ranked frequency bins. The y-axis plots the total number of syllable tokens extracted for each frequency bin.

[2] Note that recall and precision coincide since the number of boundaries (one less than the number of vowels) is constant for different segmentations of the same word.

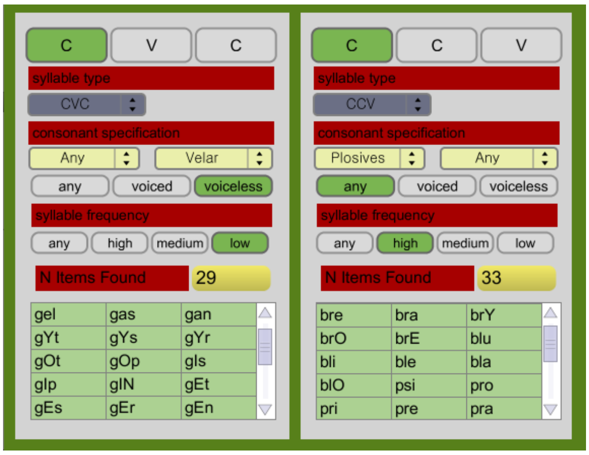
[3] A manual evaluation of the disagreement set revealed a clear advantage for the phoneme-based method

5 An interface for clinicians

In order to make the collected linguistic information available for clinicians, an interface has been built which enables clinicians to compose individual training programs. A training program consists of several training sessions, which in turn consists of a number of exercises. For each exercise, a number of syllable sets are selected, according to the specific needs of the patient. The main function of the interface, thus, deals with selection of customized syllable sets, and is described next. The rest of the interface deals with the different ways in which the syllable sets can be grouped into exercises, and how exercises are scheduled between treatment sessions.

5.1 User-defined syllable sets

The process starts with selecting the number of syllables in the current set, a number between one and four. Consequently, the selected number of 'syllable boxes' appear on the screen. Each box allows for a separate configuration of one syllable group. As can be seen in Figure 2, a syllable box contains a number of menus, and a text grid at the bottom of the box.



syllable type: CVC
consonant specification: Any, Velar, voiceless
syllable frequency: low
N Items Found 29

gel	gas	gan
gYt	gYs	gYr
gOt	gOp	gls
glp	glN	gEt
gEs	gEr	gEn

syllable type: CCV
consonant specification: Plosives, Any, any
syllable frequency: high
N Items Found 33

bre	bra	brY
brO	brE	blu
bli	ble	bla
blO	psi	pro
pri	pre	pra

Figure 2: A snapshot of the part of the interface allowing configuration of syllable sets

Here follows the list of the parameters which the user can manipulate, and their possible values:

- Syllable Type[4]
- Syllable Frequency[5]
- Voiced - Unvoiced consonant [6]
- Manner of articulation[7]
- Place of articulation[8]

Once the user selects a syllable type, he/she can further specify each consonant within that syllable type in terms of voiced/unvoiced segment choice and manner and place of articulation. For the sake of simplicity, syllable frequency ranks have been divided in three rank groups. Alternatively, the user can bypass this criterion by selecting 'any'. As the user selects the parameters which define the desired syllable type, the text grid is continuously filled with the list of syllables satisfying these criteria, and a counter shows the number of syllables currently in the grid.
Once the configuration process is accomplished, the syllables which 'survived' the selection will constitute the speech items of the current exercise, and the user proceeds to select how the syllable sets should be grouped, scheduled and so on.

6 Final remarks

6.1 Future directions

A formal usability study is needed in order to establish the degree of utility and satisfaction with the interface. One question which demands investigation is the degrees of choice that the selection tool should provide. With too many variables and hinges of choice, the configuration process for each patient might become complicated and time consuming. Therefore, a usability study should provide guidelines for an optimal design of the interface, so that its utility for clinicians is maximized.
Furthermore, we plan to integrate the proposed interface within an computer-based interactive platform for speech therapy. A seamless integration of a speech items selection module within biofeedback games for performing exercises with these items seems straight forward, as the selected items can be directly embedded (e.g., as text symbols or more abstract shapes) in the graphical environment where the exercises take place.

[4] CV, CVC, CCV, CCVC, etc.

[5] Syllables are divided in three rank groups - high, medium, and low frequency.

[6] when applicable

[7] for a specific consonant. Plosives, Fricatives, Sonorants

[8] for a specific consonant. Bilabial, Labio-Dental, Alveolar, Post-Alveolar, Palatal, Velar, Uvular, Glottal

Acknowledgments
This research is supported with the 'Mosaic' grant from The Netherlands Organisation for Scientific Research (NWO). The authors are grateful for the anonymous reviewers for their constructive feedback.

References

Aichert, I., Ziegler, W. 2004. *Syllable frequency and syllable structure in apraxia of speech.* Brain and Language, 88, 148-159.

Aichert, I., Marquardt, C., Ziegler, W. 2005. *Frequenzen sublexikalischer Einheiten des Deutschen: CELEX-basierte Datenbanken.* Neurolinguistik, 19, 55-81

Baayen R.H., Piepenbrock R. and Gulikers L. 1996. *CELEX2. Linguistic Data Consortium, Philadelphia.*

Bartlett, S. 2007. *Discriminative approach to automatic syllabication.* Masters thesis, Departmentof-Computing Science, University of Alberta.

Duffy, J.R 2005. *Motor speech disorder: Substrates, Differential Diagnosis, and Management.* (2nd Ed.) 507-524. St. Louis, MO: Elsevier Mosby

Greenberg, S. 1999. *Speaking in shorthanda syllable-centric perspective for understanding pronunciation variation.* Speech Comm., 29(2-4):159-176

Kent, R. 2000. *Research on speech motor control and its disorders, a review and prospectives.* Speech Comm., 29(2-4):159-176 J.

Kleinow, J., Smith, A. 2000. *Inuences of length and syntactic complexity on the speech motor stability of the uent speech of adults who stutter.* Journal of Speech, Language, and Hearing Research, 43, 548559.

Laganaro, M. 2008. *Is there a syllable frequency effect in aphasia or in apraxia of speech or both?* Aphasiology, Volume 22, Number 11, November 2008 , pp. 1191-1200(10)

Maas, E., Robin, D.A., Austermann Hula, S.N., Freedman, S.E., Wulf, G., Ballard, K.J., Schmidt, R.A. 2008. *Principles of Motor Learning in Treatment of Motor Speech Disorders* American Journal of Speech-Language Pathology, 17, 277-298.

Mitterer, H. 2008. *How are words reduced in spontaneous speech?* In A. Botinis (Ed.), Proceedings of the ISCA Tutorial and Research Workshop on Experimental Linguistics (pages 165-168). University of Athens.

Namasivayam, A.K., van Lieshout, P. 2008. *Investigating speech motor practice and learning in people who stutter* Journal of Fluency Disorders 33 (2008) 3251

Schiller, N. O., Meyer, A. S., Baayen, R. H., Levelt, W. J. M. 1996. *A Comparison of Lexeme and Speech Syllables in Dutch.* Journal of Quantitative Linguistics, 3, 8-28.

Schmid H., Möbius B. and Weidenkaff J. 2007. *Tagging Syllable Boundaries With Joint N-Gram Models.* Proceedings of Interspeech-2007 (Antwerpen), pages 2857-2860.

Smits-Bandstra, S., DeNil, L. F., Saint-Cyr, J. 2006. *Speech and non-speech sequence skill learning in adults who stutter.* Journal of Fluency Disorders, 31,116136.

Staiger, A., Ziegler, W. 2008. *Syllable frequency and syllable structure in the spontaneous speech production of patients with apraxia of speech.* Aphasiology, Volume 22, Number 11, November 2008 , pp. 1201-1215(15)

Tjaden, K. 2000. *Exploration of a treatment technique for prosodic disturbance following stroke training.* Clinical Linguistics and Phonetics 2000, Vol. 14, No. 8, Pages 619-641

Riley, J., Riley, G. 1995. *Speech motor improvement program for children who stutter.* In C.W. Starkweather, H.F.M. Peters (Eds.), Stuttering (pp.269-272) New York: Elsevier

Van Lieshout, P. H. H. M. 2001. *Recent developments in studies of speech motor control in stuttering.* In B. Maassen, W. Hulstijn, R. D. Kent, H. F. M. Peters, P. H. H. M. Van Lieshout (Eds.), Speech motor control in normal and disordered speech(pp. 286290). Nijmegen, The Netherlands:Vantilt.

Ziegler W. 2009. *Modelling the architecture of phonetic plans: Evidence from apraxia of speech.* Language and Cognitive Processes 24, 631 - 661

Ziegler W. 2008. *Apraxia of speech.* In: Goldenberg G, Miller B (Eds.), Handbook of Clinical Neurology, Vol. 88 (3rd series), pp. 269 - 285. Elsevier. London

Zue, V.W. and Seneff, S. 1996. *"Transcription and alignment of the TIMIT database.* In Recent Research Towards Advanced Man-Machine Interface Through Spoken Language. H. Fujisaki (ed.), Amsterdam: Elsevier, 1996, pp. 515-525.