Toward a Totally Unsupervised, Language-Independent Method for the Syllabification of Written Texts

Thomas Mayer
Department of Linguistics
University of Konstanz, Germany
thomas.mayer@uni-konstanz.de

Abstract

Unsupervised algorithms for the induction of linguistic knowledge should at best require as few basic assumptions as possible and at the same time in principle yield good results for any language. However, most of the time such algorithms are only tested on a few (closely related) languages. In this paper, an approach is presented that takes into account typological knowledge in order to induce syllabic divisions in a fully automatic manner based on reasonably-sized written texts. Our approach is able to account for syllable structures of languages where other approaches would fail, thereby raising the question whether computational methods can really be claimed to be language-universal when they are not tested on the variety of structures that are found in the languages of the world.

1 Introduction

Many approaches developed in the field of computational linguistics are only tested and optimized for one language (mostly English) or a small set of closely related languages, but at the same time are often claimed to be applicable to any natural language, cf. Bender (2009). Our aim is to stress the importance of having a more varied sample of languages that include the different types that can be found in the languages of the world in order to do justice to the range of variation in linguistic structures across languages. Furthermore, we want to point to the usefulness of using typological knowledge for a language-universal approach.

In this paper, we present an unsupervised, language-independent syllabification method based on raw unannotated texts in a phonemic transcription. The methods and procedures presented in this work rest upon insights from typological work and do not need any additional language-dependent information. The main purpose of this paper is not to present an improvement on already established statistical approaches to the problem of syllabification of an individual language, but to introduce data from languages that might constitute a problem for many syllabification methods that have been optimized on languages like English and therefore make it necessary to integrate an additional component that is able to handle such cases.

The remainder of the paper is organized as follows. First, it is argued in Section 2 that orthographic texts (in any alphabetic script) can be used for the induction of phonological patterns if the spelling system is reasonably close to a phonemic transcription. The syllabification process can be divided into two steps. In Section 3, we present and evaluate an algorithm for an unsupervised classification of all symbols in the input texts into vowels and consonants. Based on this classification, a syllabification procedure is discussed that makes use of distributional information of clusters in order to break up vowel and consonant sequences into syllables (Section 4). Finally, we conclude with a discussion of the advantages and disadvantages of the present approach and its implications for future research.

2 Learning phonological patterns on the basis of written texts?

Most studies that are based on original texts are concerned with research questions that do not make use of phonological knowledge that has been extracted from the texts. The reason for this is obvious. The orthographies of many well-studied modern languages contain many idiosyncratic rules and exceptions that would make it difficult to use them for dealing with phonological aspects of the languages under consideration. On

Proceedings of the 11th Meeting of the ACL-SIGMORPHON, ACL 2010, pages 63–71,
Uppsala, Sweden, 15 July 2010.

the other hand, in order to be able to use distributional information for phonological problems there are not enough reasonably-sized phonetically transcribed corpora, especially for a wider range of languages.

However, many spelling systems do not suffer from these shortcomings and thus can be used for these purposes. When looking at languages whose orthographies have been conceived or standardized only recently it can be noted that many of them are pretty close to a phonemic transcription. Provided the size of the corpus is big enough, smaller inconsistencies in the spelling system can be considered to be noise in the data.

Phonemic orthographies as they are usually devised for a new spelling system also show an advantage that phonetic transcriptions lack, namely that they already group together those symbols that represent the same phoneme in the language.[1] Moreover, obligatory phonological processes such as final devoicing are mostly not represented in the written form (Turkish being a notable exception), thereby providing a sort of underlying representation that is useful to induce which sequences can be grouped together to morphemes.

For these reasons written texts can in our view also be used for the induction of phonological knowledge for languages with phonemic spelling systems, even though their results have to be analyzed with great care.

3 Sukhotin's algorithm

Sukhotin's algorithm (Sukhotin, 1962, 1973) is a totally unsupervised method to discriminate vowels from consonants on the basis of a phonemic transcription. The approach relies on two fundamental assumptions that are grounded on typological insights. First, vowels and consonants in words tend to alternate rather than being grouped together. Second, the most frequent symbol in the corpus is a vowel. The latter assumption is used to initialize the classification step by claiming that the most frequent symbol is the first member of the vowel class, with the rest of the symbols initially all classified as consonants. With the help of the first assumption the other vowels are then classified by iteratively checking which symbol is less frequently adjacent to the already detected vowels.

3.1 Typological basis

It has been noticed in the typological literature at least since Jakobson and Halle (1956) that there is a tendency in the languages of the world for having CV as the basic syllable structure. Of course, languages differ as to the number and types of syllables; there are languages that allow a huge variety of consonant (or vowel) clusters whereas others are stricter in their phonotactic possibilities. However, all languages seem to obey the universal law that CV is more basic than other syllable types and that "CV is the only universal model of the syllable." Evidence for this comes from different areas of linguistics, including the observation that no matter how small the number of syllable types in a language is, it always includes CV. This is also reflected in the Onset Maximization Principle (OMP), which states that an intervocalic consonant is attributed to the following syllable and is assumed to be a language-universal principle for syllabification.

We are not aware of any cross-linguistic study that investigated the token frequency of phonemes in larger samples of texts. Hence, the second assumption that the most frequent symbol in a text is always a vowel cannot be backed up by typological knowledge. However, this claim can be supported indirectly. In his study on consonant-vowel ratios in 563 languages, Maddieson (2008) states that the ratio ranges between 1.11 and 29. The lowest value has been calculated for the isolate language Andoke, which has 10 consonants and 9 vowels. The mean value is 4.25, though. Provided that it is always the case that languages have more consonants than vowel types, it can be argued that the fewer vowels have higher token frequencies in order to be able to contribute their share to the make-up of syllables.[2] Yet this generalization is untested and could be wrong for some languages (or rather corpora of those languages). In our sample of texts in different languages, nevertheless the most frequent symbol is always a vowel.

3.2 Description of the algorithm

Sukhotin's algorithm is computationally simple and can even be illustrated with a small toy cor-

[1] In the remainder of this paper we will use the term 'symbol' as a more neutral expression for all letters in the written texts in order not to be explicit whether the spelling system really lives up to the goal of representing phonemes by letters.

[2] In the French corpus that Goldsmith and Xanthos (2009) used in their studies, the most frequent phoneme turned out to be a consonant. However, the rest of the classification was not affected and all remaining phonemes were labelled correctly.

pus.[3] Given a corpus with the inventory of n symbols $S := \{s_1, \ldots, s_n\}$ we construct an $n \times n$ matrix M where the rows represent the first and the columns the second symbol in a bigram sequence and which indicates the number of times the sequences occur in the corpus.

$$M = \begin{pmatrix} m_{11} & \ldots & m_{1n} \\ \ldots & \ldots & \ldots \\ m_{n1} & \ldots & m_{nn} \end{pmatrix}$$

The main diagonal, i.e., the self-succession of symbols, is ignored by setting all its values to zero. For instance, given a sample corpus $C = \{saat, salat, tal, last, stall, lese, seele\}$ we obtain the following 5×5 matrix (for ease of understanding the symbols have been put in front of the cells of the matrix and the row sums in the last column):

$$M = \begin{pmatrix} & s & a & t & l & e & \text{Sum} \\ s & 0 & 3 & 2 & 0 & 3 & 8 \\ a & 3 & 0 & 3 & 4 & 0 & 10 \\ t & 2 & 3 & 0 & 0 & 2 & 7 \\ l & 0 & 4 & 0 & 0 & 3 & 7 \\ e & 3 & 0 & 2 & 3 & 0 & 8 \end{pmatrix}$$

Sukhotin's algorithm initially considers all symbols to be consonants before it enters an interative phase. In each cycle of the phase, the symbol with the highest row sum greater than zero is detected and classified as a vowel. The row sum for any symbol s_a is calculated by adding up all occurrences of the symbol s_a as a first or second member in a sequence $\sum_{i=1}^{n} m_{ai}$. After a new vowel has been detected, its row sum is set to zero and all other row sums are updated by subtracting from the sum of the row of each remaining symbol twice the number of times it occurs next to the new-found vowel. This process is repeated until no more symbols with positive row sums are left. In our example, the vectors of row sums ($RSum$) for all symbols in the individual steps of the iteration phase look as follows:

$$RSum_1 = \begin{pmatrix} s & a & t & l & e \\ 8 & 10 & 7 & 7 & 8 \end{pmatrix}$$

$$RSum_2 = \begin{pmatrix} s & a & t & l & e \\ 2 & 0 & 1 & -1 & 8 \end{pmatrix}$$

$$RSum_3 = \begin{pmatrix} s & a & t & l & e \\ -4 & 0 & -3 & -7 & 0 \end{pmatrix}$$

The rationale behind this algorithm with respect to its basic assumptions is as follows. The fact that initially the symbol with the highest sum is considered to be a vowel reflects the idea that the most frequent symbol in the corpus has to be a vowel. What the row sums after each step actually contain is the difference between the number of times a symbol is found next to a consonant and the number of times it is found next to a vowel. Whenever a new vowel has been detected all occurrences of this vowel have to be subtracted from the other symbols because this symbol is no longer considered to be a consonant.

3.3 Evaluation

To the best of our knowledge, the algorithm has never been tested on a larger cross-linguistic sample. There are results for a number of languages in Sukhotin's original papers, in Sassoon (1992) and in Goldsmith and Xanthos (2009), yet almost all languages in those samples belong to the Indo-European family (except for Georgian, Hungarian and Finnish) or do not fulfill the criterion of a phonemic transcription (Hebrew). It therefore still needs to be tested on a more cross-linguistic sample of languages. In particular, it is an interesting question to see if the algorithm works even for those languages that are notorious for having many consonant clusters. On the basis of his sample of five languages, Sassoon (1992) comes to the conclusion that it works very well on those languages that have only few consonant clusters but has problems when more complex clusters are involved. However, he also notices that this effect disappears with larger text samples. Table 1 provides an evaluation of Sukhotin's algorithm on the basis of Bible texts (NT) in our sample of 39 languages. The size of the corpora in Sassoon's sample range from 1641 to 3781 characters while the Bible texts contain more than 100,000 characters (e.g., English has 716,301 characters). On average, Sukhotin's algorithm classifies 95.66% of the symbols correctly. However, this percentage also includes those languages which do not fulfill the criterion of having a suitable phonemic writing system (e.g., Russian, English, German, French). When looking only at those languages whose spelling systems are close to a phonemic transcription (or where the digraphs have been sub-

[3] More detailed descriptions can be found in Guy (1991) and Goldsmith and Xanthos (2009).

stituted by single symbols), the results are even better.

Misclassified symbols are either very infrequent and happen to occur next to symbols of the same class or are part of one of the digraphs used in the spelling system of the language. In the Maltese case, the symbol *î* is classified as a consonant because it only occurs twice in the corpus in the word *eloî* where it stands next to a symbol that is clearly a vowel. For some languages, minor modifications to the original texts have been made in order to replace the most frequent digraphs. In Swahili, for instance, with the official orthography the symbol *c* is classified as a vowel because it only occurs in the digraph *ch*. After the digraph has been replaced by a single symbol, the classification is correct in all cases. Sometimes a symbol (e.g., *h* in Warlpiri) is misclassified because it does not occur in the writing system of the language but is part of a digraph in foreign words (mostly proper names of people or locations in the Bible texts). Another problem of the approach is with orthographies that use the same symbol for both vowels and consonants. Since the classification is global, symbols like English *y*, which is a consonant in *yoghurt* and a vowel in *lady*, are always treated as either a vowel or a consonant for the whole language independent of the context where they occur. Therefore symbols in the input text should always be able to be classified to one or the other category.

As the discussion of misclassified symbols shows, the main errors in the results are not due to the algorithm itself, but a problem of the spelling systems of the texts at hand. Our results confirm the findings of Sassoon (1992) that the algorithm is sensitive to the corpus size and the frequency of occurrence of individual symbols. Larger corpora, such as Bible texts, yield much better results for these languages. Even those languages with many and very complex consonant clusters (e.g., Georgian, Croatian and Czech) get an almost perfect classification. It is remarkable that the overall distribution of the symbols makes up for those cases where consonants frequently occur in clusters. Experiments with smaller corpus sizes also revealed that one of the first symbols that get wrongly classified is the sibilant *s*. This might be another indicator for the exceptional status of sibilants with respect to syllabification and their occurrence in consonant sequences where they can violate the sonority principle (e.g., in the sequence *str* in words like *string* the consonant *s* is to the left of the consonant *t* although higher in sonority).

4 Unsupervised syllabification

Based the classification of input symbols into vowels and consonants, the syllabification procedure can then be applied. Knowledge of syllable structure is not only relevant for a better understanding of the procedures and representations that are involved in both computer and human language learning but also interesting from an engineering standpoint, e.g., for the correct pronunciation of unknown words in text-to-speech systems or as an intermediate step for morphology induction.

Several methods have been proposed in the literature for an unsupervised language-independent syllabification (see Vogel, 1977 for an overview and Goldsmith and Larson, 1990 for an implementation of a more recent approach based on the sonority hierarchy). Some methods that have been suggested in the literature (going back to Herodotus, who observed this for Ancient Greek; cf. Kuryłowicz, 1948) rely on the observation that word-medial tautosyllabic consonant clusters mostly constitute a subset of word-peripheral clusters. Intervocalic consonant clusters can therefore be divided up into a word-final and word-initial cluster. Theoretically, two types of problems can be encountered. First, those where more than one division is possible and second, those in which no division is possible.

Several approaches have been suggested to resolve the first problem, i.e., word-medial consonant sequences where there are several possible divisions based on the occurrence of word-initial and word-final clusters. O'Connor and Trim (1953) and Arnold (1956) suggest that in cases of ambiguous word-medial clusters the preference for one syllable division over another can be determined by the frequency of occurrence of different types of word-initial and word-final clusters. For this purpose, they determine the frequency of occurrence of word-initial and word-final CV, VC, etc. syllable patterns. Based on these frequencies they calculate the probabilities of dividing a word-medial sequence by summing up the values established for the different word-peripheral syllable types. The candidate syllabification with the highest sum is then chosen as the optimal division.

The approach taken here is a slight modification of the proposal in O'Connor and Trim (1953) and

Language	Vowels	Consonants
Afrikaans	a c* e i o u y á é ê í ò ó ô ú	b d f g h j k l m n p q r s t v w x ä* ë* ï* ü*
Albanian	a e g* h* i o u y ç* é ë	b c d f j k l m n p q r s t v x z
Armenian (transl.)	a e e' y' i o ch* o'	b g d z t' jh l x c' k h d' gh tw m y n sh p j r' s v t r c w p' q f
Basque	a e i o u v* á æ é í ó ö	b c d f g h k l m n p q r s t x y z à* ä* ç è* ü*
Breton	a c* e i o u ê	b d f g h j k l m n p r s t v w y z ñ ù* ü*
Chamorro	a e i o u á â é í ó ú	b c d f g h j l m n p q r s t v x y ã* ñ ü*
Croatian	a e i o u	b c d f g h j k l m n p r s t v z ä* ð ó* ć č đ š ž
Czech	a e i o u y á é í ó ú ý ě ů	b c d f g h j k l m n p q r s t v x z č ď ň ř š ť ž
Danish	a e i o u y å æ ø	b c d f g h j k l m n p r s t v x z
Dutch	a c* e i o u y	b d f g h j k l m n p q r s t v w x z
English	a e g* i o t* u	b c d f h j k l m n p q r s v w x y z
Finnish	a e i o u y ä ö	b c d f g h j k l m n p q r s t v x z
French	a e i o u à â è é ê î ô û	b c d f g h j k l m n p q r s t v x y z ç ë* ï* ù* ü* œ*
Georgian (transl.)	a e i o u h*	b g d v z t k' l m n p' zh r s t' p k gh q sh ch ts dz ts' ch' kh j
German	a e h* i o p* u y ä ö ü	b c d f g j k l m n q r s t v w x z ß
Gothic	a e i o u v* x* û	b d f g h j k l m n p q r s t w z ï* þ
Greek	α ε η ι ο υ ω	β γ δ ζ θ κ λ μ ν ξ π ρ ς σ τ φ χ ψ
Hungarian	a c* e i o u y á é í ó õ ö ú û ü	b d f g h j k l m n p r s t v x z
Icelandic	a e i o u y á æ é í ó ö ú ý	b d f g h j k l m n p r s t v x ð þ
Italian	a e h* i o u à è ì ò ù	b c d f g j k l m n p q r s t v z é*
Latin	a e i o u y	b c d f g h l m n p q r s t v x z
Maltese (rev.)	a e g* i o u à â è ì í ò ù	b d f h j k l m n p q r s t v w x z î* ċ ġ ħ ż
Mandarin (toneless)	a e i o u	ng zh ch b c d f g h j k l m n p q r s t w x y z sh
Maori (rev.)	a e i o u	g h k m n p r t ng w wh
Norwegian (Bokmål)	a e i o u y å æ é ó ø	b c d f g h j k l m n p r s t v z ë*
Potawatomi (rev.)	a e i o u	c d g k l m n p s t w sh y
Romanian	a e i o u î ă	b c d f g h j l m n p r s t v x z ş ţ
Russian	а е и о у ы (ь*) э я	б в г д ж з й к л м н п р с т ф х ц ч ш щ (ъ*) (ю*)
Scots Gaelic	a h* i o u	b c d e* f g l m n p r s t
Spanish	a e i o u á â é ê í î ó ô ù ú	b c d f g h j l m n p q r s t v x y z ñ ü*
Swahili (rev.)	a e i o u	b d f g h j k l m n p r s t v w x y z
Swedish	a e i o u y ä å é ö	b c d f g h j k l m n p r s t v x
Tagalog (rev.)	a e i o u	ng b c d f g h j k l m n p q r s t v w x y z
Turkish	a e i o u â å î ö û ü ı	b c d f g h j k l m n p r s t v y z ç ğ ş
Ukrainian	і а е и о у ь ю я є і ї	с у б в г д ж з й к л м н п р с т ф х ц ч ш щ ґ
Uma	a e g* i o u	b c d f h j k l m n p r s t w y z ï*
Warlpiri (rev.)	a e h* i o u	c f j k l m n p q r s t v w x y z
Wolof	a e i o u à é ë ó	b c d f g j k l m n p q r s t w x y ñ ŋ
Xhosa	a e g* i o t* u â	b c d f h j k l m n p q r s v w x y z

Table 1: Results for Sukhotin's algorithm on Bible texts in 39 languages. All symbols of the input Bible texts for the respective languages are listed even if they are very infrequent. For those languages marked as revised the most frequent digraphs have been replaced by a single symbol. Wrongly classified symbols are marked with an asterisk. Languages with spelling systems which notoriously contain many idiosyncratic rules are shaded. We decided to include them as a reference where the problems occur with these systems.

Arnold (1956). Instead of counting the frequency of occurrence of syllable types, the actual syllables are counted in order to determine the best split of word-medial consonant sequences. An example calculation for the German word *fasten* 'to abstain from food' is given in Table 2.

a)	fa $st_{142]}$ en	[fast.en]	sum: 142
b)	fa $s_{528]}$ $_{[216}$t en	[fas.ten]	sum: 744
c)	fa $_{[176}$st en	[fa.sten]	sum: 176

Table 2: Example calculations for the word-medial cluster in the German word *fasten*.

The example calculations in Table 2 show that the candidate syllabification in b) yields the highest sum and is therefore chosen as the correct syllabification of the word. One of the advantages of this approach (as well as the one proposed by O'Connor and Trim and Arnold) is that OMP follows from the fact that word-initial CV sequences are more frequent than word-final VC sequences and does not have to be stipulated independently.

The claim that CV is the unmarked syllable structure for all languages of the world (and OMP a universal principle) has been challenged by some Australian languages that seem to behave differently with respect to syllabification of VCV sequences (Breen and Pensalfini, 1999). In those languages VCV sequences are syllabified as VC.V instead of V.CV, as OMP would predict. The authors provide evidence from a number of processes in these languages (reduplication, language games) as well as historical and comparative evidence that support the analysis that VC seems to be more accurate as the basic syllable type for those languages.[4]

For cases where word-medial clusters cannot be broken up by sequences that are found at word edges (bad clusters), we decided to go back to the original method used by O'Connor and Trim and Arnold and calculate the frequency of occurrence of syllable types. However, bad clusters are not very frequent compared to the overall data in our experiments.

One additional problem when working with written texts[5] rather than transcribed corpora is the fact that diphthongs are not clearly distinguished from sequences of monophthongs. Yet this is vital for a correct syllabification procedure since the number of syllables of the word is different depending on this choice. In order to retrieve the diphthongs of the language from the distribution of vowel sequences in the corpus the following approach has been used.[6] For each bigram vowel sequence the number of times the first vowel v_1 is directly followed by the second vowel v_2 is compared with the number of times both vowels are separated by one consonant. If the frequency of direct adjacency is higher than the frequency of v_1cv_2 the sequence is considered to be a "diphthong"; if not, the sequence is considered to be a case of hiatus and both vowels are attributed to different syllables. Similar to Sukhotin's algorithm the present syllabification algorithm is also global in the sense that the diphthong/monophthong distinction is always used in the same way no matter in which environment the sequence occurs.[7] Table 3 gives a list of the diphthongs extracted from the corpus for a number of languages in our sample based on this method.

Dutch	aa (772), oo (510), ie (440), ui (301), ou (155), eu (110), uu (27)
German	ei (1373), au (641), eu (216)
English	ea (336), ou (280), io (231), oo (79)
French	ai (863), ou (686), eu (397), io (339), ui (272), au (232), oi (232)
Greek	ου (1687), ει (1684), ευ (650), οι (616), αυ (287)
Wolof	aa (1027), ee (821), oo (656), ée (181), ii (158), óo (118)

Table 3: "Diphthongs" for a subset of the languages in the sample (in brackets the frequency of adjacent occurrence).

4.1 The problem of evaluating syllabification methods

There are several reasons why a gold standard for syllabification, with which the syllabification methods are compared, is difficult to establish.

[4]Note that this does not invalidate one of the basic assumptions of Sukhotin's algorithm, since C and V still alternate even though in the reverse order.

[5]Some linguists also believe that stress can lead to a violation of OMP by attracting an intervocalic consonant to the coda of the previous stressed syllable. Since stress is usually not reflected in most orthographies, we do not consider this option here.

[6]We thank Bernhard Wälchli (p.c.) for drawing our attention to this idea.

[7]In German, for instance, the vowel sequence *eu* can either be tautosyllabic and in that case constitute a diphthong as in *heute* 'today'; or it can be a case of hiatus and therefore be broken up by a syllable boundary as in *Museum* 'museum'.

Duanmu (2009) states that even for well-described languages like English linguists do not agree on the correct syllabification of comparatively straightforward cases. For the English word *happy*, for instance, four different analyses have been proposed:

[hæ.pi]	Hayes (1995), Halle (1998), Gussmann (2002)
[hæp.i]	Selkirk (1982), Hammond (1999)
[hæp̲i]	Kahn (1976), Giegerich (1992), Kreidler (2004)
[hæp.pi]	Burzio (1994)

Table 4: Analyses of *happy* (cited from Duanmu, 2009). Underlined consonants are ambisyllabic.

The correct syllabification of a word can best be established when there is some operation in the language that takes recourse on the syllable structure of the word. In the case of the Australian languages with no syllable onsets, Breen and Pensalfini (1999:6f) provide evidence from reduplication processes in Arrernte to support their analysis. If the Arrernte syllable shape is VC(C), rather than (C)CV, reduplication is most straightforwardly described in terms of syllables. The attenuative prefix is formed by /-elp/ preceded by the first syllable of the base if VC(C) syllabification is assumed. The attenuative form of the base *emp^{w}aṛ* 'to make' is therefore *emp^{w}elpemp^{w}aṛ*.[8] A similar argumentation can be put forward for languages that show phonological operations that are based on the structure of syllables, e.g., syllable-final devoicing. If a voiced obstruent is realized unvoiced, the syllabification might suggest its position to be in the coda.

Besides disagreement on the correct syllabification of words, another crucial aspect of evaluating syllabification methods is the question of whether the test set should consist of a random sample of words of the language or whether there should be any constraints on the composition of the evaluation data. If the evaluation consists of a huge number of monosyllabic words, the results are much better than with polysyllabic words because no consonant clusters have to be broken up.

[8] As one reviewer remarked, reduplication patterns are usually described in terms of a CV-template rather than syllable structures. However, in the case of Arrernte, a description in terms of syllables rather than VC(C) shapes would be more elegant and at the same time account for other operations as well.

For the evaluation of their syllabification methods, Goldwater and Johnson (2005) distinguish words with any number of syllables from words with at least two syllables. Depending on the method that they test the differences in the percentage of correctly syllabified words range from a few to almost 30%. It is therefore easier to get better results when applying the syllabification methods to languages with a large number of monosyllabic words and fewer consonant clusters, like Mandarin Chinese, for instance.

4.2 Discussion and evaluation

One of the problems of a cross-linguistic investigation is the availability of gold standards for evaluation. Thus, instead of providing a comparative evaluation, we want to discuss the advantages and disadvantages of the procedure with respect to the more common sonority-based syllabification method. We tested our method on a manually created gold standard of 1,000 randomly selected words in Latin. The precision is 92.50% and the recall 94.96% (F-Score 0.94) for each transition from one symbol to another. Most misplaced syllable boundaries are due to the vowel cluster *io*, which has been treated as a diphthong by our method.

The most interesting aspect of our approach is that it is able to account for those languages where intervocalic consonants are better be analyzed as belonging to the previous syllable, thereby violating OMP. Approaches relying on the Onset Maximization Principle would get all of these syllable boundaries wrong. Breen and Pensalfini (1999) note that Arrernte also has only VC in word-initial position. Consequently, an approach that is based on word-peripheral clusters can predict the lack of word-medial onsets correctly. The importance of word-peripheral clusters is also supported by findings in Goldwater and Johnson (2005) where a bigram model improves after training with Expectation Maximization whereas a positional model does not, which might be due to the fact that a bigram model (unlike the positional model) can generalize whatever it learns about clusters no matter if they occur at word edges or word-medially.

Moreover, the influence of word-peripheral clusters on the syllabification of word-medial consonant sequences is not restricted to syllable types only, but sometimes also holds solely for individual consonants. In Chamorro, for instance, Topping (1973) describes the syllabification of intervocalic consonants as observing OMP. However,

this does not apply if the consonant is the glottal stop /'/, in which case the syllable division occurs after the consonant, leading to the syllabification /na'.i/ 'to give'. The interesting observation in this respect is that the glottal stop phonologically never occurs at the beginning of a word in Chamorro whereas all other consonants (with the exception of /w/) do occur word-initially,[9] which leads to the correct syllabification results with our approach.

Another advantage of the present method is that clusters with sibilant consonants that do not conform to the sonority principle (see the example of *str* in Section 3.3) do not have to be treated differently. They merely follow from the fact that these clusters are particularly frequent in word-peripheral position. The biggest disadvantage is the fact that the method is sensitive to frequencies of individual clusters and thereby sometimes breaks up clusters that should be tautosyllabic (one of the few examples in our Latin corpus was *teneb.rae*).

5 Conclusions and future work

A complete model of syllabification involves more than what has been presented in this paper. The method proposed here is restricted to single words and does not take into account resyllabification across word boundaries as well as some other criteria that might influence the actual syllable structure of words such as stress and morphological boundaries. Nevertheless, the discussion of our approach shows that expanding the range of languages to other families and areas of the world can challenge some of the well-established findings that are used for inferring linguistic knowledge.

The results of Sukhotin's algorithm show that the distinction between vowels and consonants, which is vital for any syllabification method, can be induced from raw texts on the basis of the simple assumptions that vowels and consonants tend to alternate and that a vowel is the most frequent symbol in a corpus. In contrast to previous studies of the algorithm (Sassoon, 1992), our results do not suffer from the fact that the input text is too short and therefore yield better results.

Based on the classifications of symbols into vowels and consonants with Sukhotin's algorithm our unsupervised syllabification method determines syllable boundaries on distributional information. In contrast to other unsupervised approaches to syllabification that are grounded on attributing a sonority value to each consonant and OMP, our procedure breaks up word-medial consonant sequences by considering the frequencies of all possible word-peripheral clusters in order to get the most probable division. We did not provide a comparative evaluation of our procedure but only discussed the problems that can be encountered when looking at a wider variety of languages and how they can be solved by our approach. The question that this paper wants to raise is therefore if it is more important to optimize a procedure on a single language (mostly English or related European languages) or whether it should be capable of dealing with the variety of structures that can be found in the languages of the world.

For future work we want to apply the present methods on phonetically transcribed corpora in order to be able to compare the results for the well-studied European languages to other methods. There are still some challenges remaining for a universal syllabification procedure, one of them being the detection of syllabic consonants. Ultimately, we also want to integrate a sonority hierarchy of the input symbols to combine the advantages of both approaches and to create a gradual value for syllabification that is able to account for the difference between clear-cut syllable boundaries and ambisyllabic consonants or other cases where a syllable boundary is harder to establish.

[9] Topping notes that phonetically there is a glottal stop preceding every word-initial vowel, yet this is totally predictable in this position and therefore not phonemic.

Acknowledgments

This work has been funded by the research initiative "Computational Analysis of Linguistic Development" and the DFG Sonderforschungsbereich 471 "Variation und Entwicklung im Lexikon" at the University of Konstanz. The author would like to thank the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) for the Warlpiri Bible sections as well as Miriam Butt, Frans Plank, Bernhard Wälchli and three anonymous reviewers for valuable comments and suggestions.

References

Gordon F. Arnold. 1955-1956. A phonological approach to vowel, consonant and syllable in modern french. *Lingua*, V:251–287.

Emily Bender. 2009. Linguistically naive != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*, pages 26–32.

Gavan Breen and Rob Pensalfini. 1999. Arrernte: A language with no syllable onsets. *Linguistic Inquiry*, 30(1):1–25.

Luigi Burzio. 1994. *Principles of English Stress*. Cambridge: Cambridge University Press.

San Duanmu. 2009. *Syllable Structure*. Oxford: Oxford University Press.

Heinz Giegerich. 1992. *English Phonology*. Cambridge: Cambridge University Press.

John Goldsmith and Gary Larson. 1990. Local modelling and syllabification. In Michael Ziolkowski, Manuela Noske, and Karen Deaton, editors, *The Parasession on the Syllable in Phonetics & Phonology*, volume 2 of *Papers from the 26th Regional Meeting of the Chicago Linguistic*, pages 130–141. Chicago Linguistic Society.

John Goldsmith and Aris Xanthos. 2009. Learning phonological categories. *Language*, 85(1):4–38.

Sharon Goldwater and Mark Johnson. 2005. Representational bias in unsupervised learning of syllable structure. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CONLL), Ann Arbor*.

Edmund Gussmann. 2002. *Phonoloy: Analysis and Theory*. Cambridge: Cambridge University Press.

Jacques B. M. Guy. 1991. Vowel identification: an old (but good) algorithm. *Cryptologia*, XV(3):258–262, July.

Morris Halle. 1998. The stress of english words. *Linguistic Inquiry*, 29(4):539–568.

Michael Hammond. 1999. *The Phonology of English: A Prosodic Optimality Theoretic Approach*. Oxford: Oxford University Press.

Bruce Hayes. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicage: University of Chicage Press.

Roman Jakobson and Morris Halle. 1956. *Fundamentals of Language I. Phonology and Phonetics*. 's-Gravenhage: Mouton.

Daniel Kahn. 1976. *Syllable-based generalizations in English phonology*. Ph.D. thesis, Massachusetts Institute of Technology.

Charles W. Kreidler. 2004. *The Pronunciation of English: A Course Book*. Malden, MA: Blackwell.

Jerzy Kuryłowicz. 1948. Contribution à la théorie de la syllabe. *Bulletin de la Societe Polonaise de Linguistique*, 8:5–114.

Ian Maddieson. 2008. Consonant-vowel ratio. In Martin Haspelmath, Matthew S. Dryer, David Gil, and Bernard Comrie, editors, *The World Atlas of Language Structures Online*, chapter 3. Munich: Max Planck Digital Library. Available online at http://wals.info/feature/3. Accessed on 2010-04-23.

J. D. O'Connor and J. L. M. Trim. 1953. Vowel, consonant, and syllable - a phonological definition. *Word*, 9(2):103–122.

George T. Sassoon. 1992. The application of Sukhotin's algorithm to certain Non-English languages. *Cryptologia*, 16(2):165–173.

Elisabeth O. Selkirk. 1982. The syllable. In Harry van der Hulst and Norval Smith, editors, *The Structure of Phonological Representations, part II*, pages 337–383. Dordrecht: Foris.

Boris V. Sukhotin. 1962. Eksperimental'noe vydelenie klassov bukv s pomoşçju evm. *Problemy strukturnoj lingvistiki*, 234:189–206.

Boris V. Sukhotin. 1973. Méthode de déchiffrage, outil de recherche en linguistique. *T.A. Informations*, 2:1–43.

Donald M. Topping. 1980. *Chamorro Reference Grammar*. The University Press of Hawaii, Honolulu.

Irene Vogel. 1977. *The Syllable in Phonological Theory with Special Reference to Italian*. Ph.D. thesis, Stanford University.