The Relation of Writing to Spoken Language

Edited by
Martin Neef, Anneke Neijt and Richard Sproat

Sonderdruck

aus Linguistische Arbeiten, Band 460
ISBN 3-484-30460-X

Max Niemeyer Verlag
Tübingen 2002

2002

Susanne R. Borgwaldt & Annette M.B. de Groot

Beyond the Rime: Measuring the Consistency of Monosyllabic and Polysyllabic Words

1. Introduction

One of the various variables that affect visual word recognition is the degree of spelling-sound (un)ambiguity. Words with predictable spelling-sound correspondences cause shorter reaction times than words with ambiguous spelling-sound correspondences. One way to describe these spelling-sound relations is to use the *'word consistency paradigm'* (Glushko 1979).

Traditionally, psycholinguistic research on phonological consistency has focused on monosyllabic words. These words are split up into onset and rime. Subsequently, the mappings between written and spoken rimes are compared. Words sharing the same written rime are considered *feedforward consistent* if the corresponding spoken rimes are pronounced in the same way. Words sharing the same spoken rime are called *feedback consistent* (Stone et al. 1997) if their rimes are written in the same way.

In this article, we present a methodology for determining the degree of bidirectional consistency that is applicable for monosyllabic and polysyllabic data alike. We compare this approach with the traditional (monosyllabic) rime analyses. We show that by taking not only the consistency mappings between rimes into account but also those between other (overlapping) subsyllabic units, we can expand the range and increase the accuracy of the description of consistency considerably.

In contrast to other orthographies like the Chinese writing system, where characters represent morphemes, or syllabaries like the Japanese hiragana and katakana, in which each symbol corresponds to a mora or to a certain syllable type respectively, alphabetic writing systems are generally based on the principle of grapheme-phoneme correspondence, where 'grapheme' refers to one single letter or a letter combination that corresponds to a single phoneme (cf. e.g. Neef, this volume). Nevertheless, the individual languages written in an alphabet deviate from this rule in various degrees. In a very regular language like Finnish, a one-to-one relation between graphemes and phonemes is almost always realized; the pronunciation of words is predictable from their spelling and vice versa (Karlsson 1983). A well-known counterexample is English, displaying ambiguous and often unpredictable relations between spelling and pronunciation.

Although grapheme-phoneme relations of a language are bidirectional in nature, psycholinguistic research has mainly focused on one direction, namely on the mappings from spelling to pronunciation. This is probably a result of the dominance of research on reading as compared to research on spelling (Brown & Ellis 1994), and of one of the views emerging from this research on reading, namely the view that word reading is only influenced by spelling-pronunciation irregularities and inconsistencies and not by irregularities and inconsisten-

cies in the other direction. However, some researchers (Stone et al. 1997, Ziegler et al. 1997) have challenged this assumption.

The degree of phonological consistency of a given language depends on the unit size analyzed. The pronunciation of letters and graphemes in isolation, or the spelling of single phonemes, is often less predictable than the spelling or pronunciation of larger sublexical units (Treiman et al. 1995, Martensen et al. 2000, Kessler & Treiman 2001). A study of the English language (Dewey 1971) finds on average 13.7 different spellings corresponding to one phoneme and 3.5 different phonemes per letter. Other analyses focusing on grapheme-phoneme correspondences also report a great number of irregularities (Venezky 1970, Berndt et al. 1987, for English; Véronis 1986, for French), whereas more systematic relations between spelling and pronunciation can be found for larger constituents (Treiman et al. 1995).

Naming studies by Treiman et al. (1995) showed that spelling-to-sound correspondences of rimes[1] of monosyllabic English words play an important role in reading, as compared to the mappings of other subsyllabic units, like onset, nucleus, coda or body. Research has therefore moved towards studying correspondence relations on higher levels, focusing on the rimes of monosyllabic words.

Though this procedure is empirically based on the English language, research comparing the consistency of rime correspondences has extended to other languages as well, for example to French (Ziegler et al. 1996), and Dutch (Martensen et al. 2000). Ziegler et al. (1996, 1997) analyzed the bidirectional inconsistencies of English and French rimes and explored the influence of ambiguous spelling-to-sound and sound-to-spelling mappings on visual word recognition performance. They discovered that in both languages not only feedforward (spelling-to-sound) inconsistent rimes but also feedback (sound-to-spelling) inconsistent rimes influence performance in lexical decision tasks.

As the consistency paradigm has been introduced for monosyllabic rimes (Glushko 1979), its strength lies in analyzing the consistency of languages with a high percentage of monomorphemic monosyllabic words like English, which is typologically characterized as a relatively isolating language. However, its applicability and the informative value of the analyses are limited for languages with a high amount of polysyllabic and polymorphemic words and word forms. For example, in Finnish, classified as an agglutinative language, almost no monosyllabic words exist (Karlsson 1983), Spanish, Portuguese, Italian and German, morphologically categorized as fusional languages, also contain relatively few monomorphemic monosyllabic words. For those languages, the computation of bidirectional rime-consistency statistics does not yield predictive results for estimations of the overall phonological consistency. First, we cannot generalize from samples that small; and, secondly, a small corpus is automatically biased towards consistency, as it contains a relatively high percentage of unique and, therefore consistent, rimes.

---

[1] There are several names for this sublexical entity. The term 'rhyme' is often used in phonetics and theoretical linguistics, mainly referring to the spoken unit. Many psycholinguists seem to opt for 'rime', denoting both spoken and written units. Other authors use the term 'body', but this might lead to confusion, as the same word is used by phonologists to describe the concatenation of onset and nucleus (Iverson & Wheeler 1989). Henceforth we will use the term 'rime' in this article, referring to the vowel and following consonant(s) of a monosyllabic word, for both written and spoken entities.

For example, an analysis of 1149 German monosyllabic words[2] yielded the following results for the spelling-to-sound mappings: only 31 written rimes, corresponding to 153 words, are feedforward inconsistent, i.e., they have more than one pronunciation. The other 408 written rimes, that correspond to 996 words, are pronounced consistently, but of those consistent rimes, 215 are unique. Analyzing the sound-to-spelling mappings showed that 99 spoken rimes, corresponding to 567 words, are feedback inconsistent, i.e., they have more than one spelling. In this analysis of sound-to-spelling mappings, 244 rimes, corresponding to 582 words, are written consistently, but among those, 130 are unique. Although it can be concluded from this sample that German is more consistent in the spelling-to-sound direction than in the reverse direction, the high percentage of unique rimes reflects a strong bias for consistency in this sample, and subsequently limits the reliability of our analysis.[3]

Note, that the terms 'consistent' and 'regular', although highly correlated, are not synonymous with each other, but concern dissociable variables. A word like *YACHT*, pronounced[4] /jɒt/, is considered to be irregular in the mapping from the written to the spoken rime, but it is nevertheless consistent, as its orthographic rime is unique: No other word exists in which the same written rime is pronounced differently.

Aiming at an extension of the lexical coverage of the analysis by applying rime comparisons also to polysyllabic words[5] gives rise to a new obstacle. While the concept 'rime' is relatively well defined for monosyllabic words (i.e. the vowel and following consonant(s)), how the rime or rimes of polysyllabic words should be defined conceptually, segmented, and subsequently analyzed, are questions yet to be answered.

A feasible procedure, analogous to the treatment of monosyllabic words, follows the criteria applied in the selection of rhyming words in poetry: the phonological rime of a polysyllabic word is the concatenation of the rime of the last stressed syllable of this word with all following syllables. In this way, the phonological rime of *PALATE* is /'ælət/, rhyming with *BALLOT*, and the corresponding orthographic rimes are <alate> and <allot> respectively. The orthographic and phonological rimes of *DESSERT* are <ert> and /'ɜ:t/ respectively, rhyming with *SHIRT*, and the rimes of *BOOMERANG* are <oomerang> and /'u:məræŋ/, rhyming with no other word: they are both unique. However, this method will yield relatively huge numbers of unique rimes (as with the increasing number of letters or phonemes in the rime, the chances decrease that other words contain exactly the same letters or phonemes in their rime) and

---

2 This sample contains all monosyllabic adjectives, nouns and verbs with a word frequency count of 1 per million or higher, extracted from the Celex database (Baayen, Piepenbrock & Gulikers 1995), a lexical database for English, Dutch and German. (Borgwaldt & De Groot, in preparation).

3 A rime analysis of a corpus of 2124 monosyllabic Dutch words, compiled according to the same criteria from the Celex database (Baayen et al. 1995), results in only 193 feedforward unique rimes and 114 feedback unique rimes.

4 The English pronunciations in this chapter follow the notation for British English in the Cambridge International Dictionary of English. However, our transcription deviates slightly from dictionary entries: first, we do not mark syllable boundaries, and second, we indicate lexical stress at the nucleus of the stressed syllable and not at the onset of the syllable.

5 The addition of polysyllabic monomorphemic words will increase, for example, the German corpus to over 5000 words, compared to only 1149 monosyllabic ones. For the Dutch Celex corpus mentioned above, the augmentation by polysyllabic words will result in a corpus of over 8000 words, compared to 2124 monosyllabic ones.

subsequently biased consistency percentages, as no deviating pronunciation or spelling can be found.

An analysis of the 3175 German monomorphemic polysyllabic words with stress falling on the penultimate syllable (see footnote 2 for selection criteria) yielded the following results for the spelling-to-sound mappings of the disyllabic rimes: only 23 written rimes, corresponding to 95 words, are feedforward inconsistent. The other 1633 written rimes, corresponding to 3080 words, are pronounced consistently, but 1085 of those are unique. Investigating the sound-to-spelling mappings showed that 88 spoken rimes, corresponding to 411 words, are feedback inconsistent; 1500 rimes, corresponding to 2764 words, are written consistently, but among those 1028 are unique. This analysis supports the results for the monosyllabic rime analysis mentioned above that German is indeed more consistent in spelling-to-sound direction than in sound-to-spelling direction, but again the high percentage of unique rimes results in a strong bias towards consistency.

A different way to examine the consistency of polysyllabic words using rime mappings is to segment these words first into syllables, then to split up each syllable in onset and rime, and finally to analyze these rimes separately. However, by choosing this approach we are, at least for some languages, faced with ambisyllabic phonemes and graphemes. In German we find consonant phonemes that can be assigned to both the coda of one syllable and the onset of the next syllable, like /m/ in *HUMMER*, pronounced /hˈʊmər/. The corresponding grapheme <mm> is ambisyllabic as well, as *HUMMER* is hyphenated like *HUM-MER*.[6] Another problematic case for analyses of separate syllables are words in German that contain the grapheme <ng>, denoting regularly the phoneme /ŋ/, as in *ENG*, pronounced /ɛŋ/, or *ENGEL*, pronounced /ˈɛŋəl/. By decomposing <engel> in its two orthographic syllables, <en> and <gel>, we cease to capture the regularity of this mapping, as we have to split up the grapheme <ng> into two separate letters, <n> and <g>, that are pronounced differently in isolation. The occurrence of these cases and others, where the boundaries of the phonological and orthographic syllables do not match, are unclear, or variable, demonstrate at least for some languages the need for an analysis of the spelling-to-sound mappings that exceeds one syllable.

Besides capturing the mappings between pronunciation and spelling of both mono- and polysyllabic words in a language by consistency measurements, it is clearly possible to accomplish this task by using rule approaches, as demonstrated by e.g. Venezky (1970), Baron et al. (1980), and Nunn (1998). However, comparing rule-based descriptions of the spelling-sound relations of individual languages will pose difficulties for cross-linguistic comparisons. By using a consistency approach, only percentages of consistent or inconsistent mappings between pronunciation and spelling patterns will have to be compared across languages. By utilizing rule-based approaches, however, we will also have to take into account that rule-systems can not only vary in the number of rules but also in terms of additional complexity features of those rules, like for example scope, cyclicity and interaction of rule sequences and number of exceptions. These features will all have to be measured and included in the comparisons as well.

There are related research domains that aim at generating knowledge about the spelling-pronunciation mappings of languages in the form of statistical models. One is the area of

---

[6] For a detailed analysis of principles of hyphenation in German, see Geilfuß-Wolfgang (this volume).

machine learning (Van den Bosch et al. 1995, Yvon 1997) in which aligned orthographic and phonological transcriptions of a corpus are moved through sliding windows composed of several phonemes or letters. Another related research area is speech synthesis. Some of the newer large vocabulary speech synthesis system components investigate pronunciation by analogy (Marchand & Damper 2000), predicting the pronunciation of novel words by using letter-phoneme mappings of substrings of varying grain sizes. Both approaches usually do not take morphemic or intrasyllabic structures into account and, therefore, produce biased results. We will discuss the implications of this in a later section.

2. Computing the phonological consistency of mono- and polysyllabic words

Our requirements for a valid method to measure phonological consistency cross-linguistically are as follows: The method should not be restricted to monosyllabic words, as the percentage of words fulfilling this criterion varies among languages. Furthermore, it should not be limited to the analysis of rime inconsistencies, as in principle also ambiguous mappings between spelling and pronunciation in other sublexical units can exist, even if they might influence visual word recognition performance less substantially.

Our approach, which integrates both the concept of sliding windows used in machine learning techniques and the traditional comparisons of the spelling-sound mappings of sublexical units, will expand the range and increase the accuracy of the analysis. It will allow us to analyze mono- and polysyllabic words and comparing them with each other, thereby reducing the number of unique rimes and other sublexical units that are unique.

The presently proposed procedure is still limited in that it only determines the degree of spelling-to-sound consistency of monomorphemic words in isolation. Therefore, from the degree of phonological consistency as measured according to this procedure, we cannot draw strong conclusions regarding the difficulty of spelling in 'real life situations', because in connected speech, the pronunciation may undergo substantial changes by processes like assimilation, liaison, epenthesis, elision, and other effects of coarticulation.

For any language whose phonological consistency we want to analyze we need a reasonable comprehensive list of words, together with their word class, morphemic status, a transcription of their canonical pronunciation and preferably frequency information.

We will prepare our corpus by extracting all monomorphemic nouns, adjectives and verbs from the database, as it will be convenient to limit our analysis to the open, productive classes. Words belonging to the closed classes (e.g. function words like prepositions) are often spelled or pronounced in a deviant way. Furthermore, it might be expedient to exclude 'non-standard' token types, such as for instance letter names, abbreviations and acronyms. Proper names and foreign loans can be marked separately, as it will be interesting for a qualitative analysis of the data at a later point to compare their influences on the overall consistency. Analogously, orthographic or pronunciation variants can be marked and analyzed at a later state.

The results of any spelling-to-sound correspondence analysis depend on the chosen transcription level of the pronunciation on the continuum from (underlying) phonemic to (surface) phonetic transcription. This is especially an issue for those languages that neutralize phonemic contrasts in their phonetic realization. Examples are flapping in American English and final obstruent devoicing in German and Dutch. In order to gain valid cross-linguistic results, it is therefore important to choose comparable levels of representation, while research on consistency within one language might profit by distinct analyses varying the levels of pronunciation representation.[7]

We will analyze only the monomorphemic vocabulary of a language, as in some languages the morphemic structure of words influences various aspects of the pronunciation and spelling of a word. By disregarding the information provided by the morphemic structure, our analysis might report a much larger number of inconsistencies than would be the case if this factor had been taken into account.

We will first describe our method for monosyllabic words. Then we will extend the analysis, mutatis mutandis, to polysyllabic words. For the sake of simplicity, most examples in this paper show the treatment of measuring feedforward consistency, i.e., relations from spelling to pronunciation. But the same method can be applied in the reverse direction, from pronunciation to spelling, in order to determine the feedback consistency.

Assuming we have a list with monosyllabic, monomorphemic nouns, adjectives and verbs as well as a transcription of their pronunciation, we first decompose the spoken syllables into their subsyllabic units. Different theories exist about the internal structure of spoken syllables, but most authors agree that a syllable can be split up into three elements: an (optional) onset, a nucleus, and a (optional) coda. The onset of a syllable consists of its initial consonant(s); the nucleus consists of the vowel or diphthong, and the coda is the final (cluster of) consonant(s). Nucleus and coda together form a unit called the rime; nucleus and onset together form a unit called the body.

For example, the syllable /pɪt/, corresponding to the word *PIT*, can be segmented into the onset /p/, the nucleus /ɪ/, and the coda /t/. The rime of /pɪt/, the concatenation of nucleus and coda, is /ɪt/, and the body of /pɪt/, the concatenation of onset and nucleus, is /pɪ/. Analogously, *SPLIT* /splɪt/ can be divided into the complex onset /spl/, the nucleus /ɪ/ and the coda /t/, whereas the word *IT* can only be decomposed into the nucleus /ɪ/ and the coda /t/, as its onset is empty. In these examples, the body of /splɪt/ is /splɪ/, and the body of /ɪt/ is the same as the nucleus, namely /ɪ/, and both words share the same rime, i.e. /ɪt/.

The next step is to segment the written syllables in the same way into onsets, nuclei and codas and aligning them with the spoken subsyllabic units. As the decomposition into sublexical units has been developed for spoken syllables, the analogous assignment of graphemes to the sublexical units of the written syllables sometimes seems arbitrary.[8]

---

[7] In this context, compare Sproat's concept of one orthographic relevant level (ORL) per language, serving as consistent representation level (Sproat 2000, this volume).

[8] Examples are words in which not all letters are pronounced, like <gh> in *NIGHT*, or the silent <e> in *MAKE*, *FILE* or *BONE*. For example, in the case of *NIGHT* we are faced with the question whether to assign the silent letters <gh> to the orthographic nucleus, resulting in a nucleus <igh>, or to the orthographic coda, resulting in the coda <ght> (for detailed discussions, see Kessler & Treiman 2001).

However, our example *PIT* can be segmented relatively easily. We decompose the spoken form /pɪt/ into the phonological onset /p/, nucleus /ɪ/ and coda /t/. Analogously, we decompose the written word *PIT* into its orthographic onset <p>, nucleus <i> and coda <t>. Then we merge onset and nucleus to form the body, and we merge nucleus and coda to form the rime. This is done for both spoken and written syllables. Subsequently, we align the spoken and written clusters. For the example *PIT*, this procedure gives us the body-pair (/pɪ/, <pi>), and the rime-pair (/ɪt/, <it>), as shown in Figure 1.

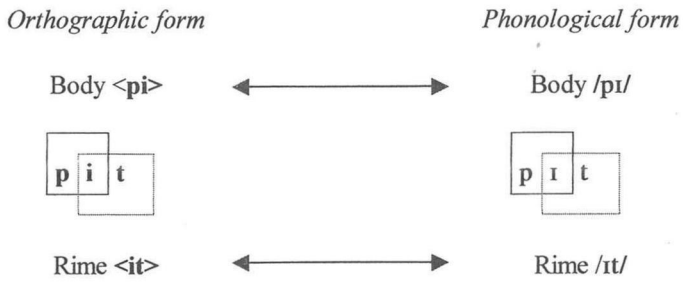

Figure 1: *Orthographic and phonological sublexical units of a monosyllabic word*

In this way, we segment all (monosyllabic, monomorphemic) words in our database. The next step is, first, to compare the mappings of the bodies with each other and then to compare the mappings of the rimes. If the mappings of the bodies are ambiguous, we check whether the mappings of the rimes may serve as disambiguating context: in our example, the body of *PIT*, <pi>, is feedforward inconsistent, as other words with the same written body but different pronunciation exist (e.g. *PINT*, /paɪnt/). In other words, when comparing only the mappings between the orthographic and phonological bodies, the pronunciation is ambiguous. However, when we take the mappings of the rime into account as well, the pronunciation is disambiguated, as the orthographic rime <it> is consistently pronounced /ɪt/.

By expanding the consistency analysis of monosyllabic words in this way, we can also take onset inconsistencies into account that are neglected by pure rime comparisons. By using overlapping units as contexts for disambiguation, we can distinguish onset inconsistencies (e.g., the different orthographic onsets in *NIGHT* vs. *KNIGHT*) from mere ambiguities of the onset in isolation. The latter ambiguous pronunciations or spellings of the onset are consistent in a larger context, in this case the body. An example is the orthographic onset <c>, ambiguous in isolation but quite predictable if the pronunciation of the bodies is analyzed (e.g., <c> will be pronounced as /k/ in words like *CAB*, *COME* and *CURT* and as /s/ in words like *CEDE*, *CITE* and *CYST*).

We analyze the polysyllabic words in our corpus in an analogous way. First we segment the word into its syllables, then into its subsyllabic units. As in some languages we might be faced with ambisyllabic segments or variable syllable boundaries, it will be convenient to merge the coda of one syllable with the onset of the next one and treat the two as one unit[9].

[9] As we are only analyzing monomorphemic words, we are not faced with morpheme boundaries determining the pronunciation of ambisyllabic consonant clusters, like in the German compound nouns *FRÜHJAHRSPUTZ*

For example, we segment the intuitively regular and consistent Dutch (loan) word *BOEMERANG* first into its three syllables. We then decompose each written and spoken syllable into the subsyllabic units onset, nucleus and coda. In our example, the phoneme /m/ is ambisyllabic and will be assigned to both coda of the first and onset of the second syllable. Finally, we merge, where applicable, the coda of one syllable with the onset of the next one. This way, for every word the following pattern will emerge: (C) V (C) V..., where V denotes the vowel or diphthong, the nucleus of a syllable, and C denotes the (optional) consonant(s) of the onset or the coda, or the cluster of the consonant(s) in the coda of one syllable and in the onset of the next one.

In Figure 2 we see the CVCVCVC pattern of *BOEMERANG*.

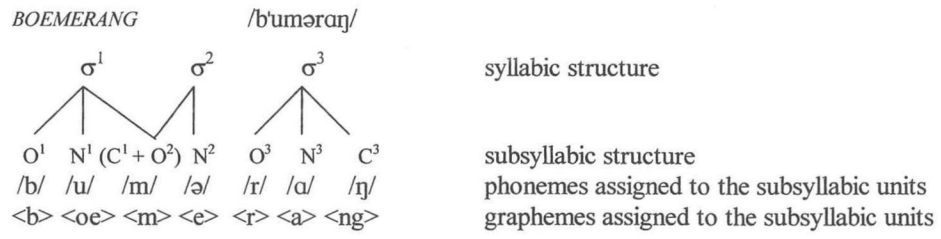

Figure 2: *Syllabic, subsyllabic and phonemic/graphemic components of a polysyllabic word*

The next step, analogous to merging the sublexical units to form the body and rime in monosyllabic words, is the construction of Overlapping SubLexical Units (henceforth referred to as OSLUs) from the subsyllabic units for polysyllabic words. To capture the body and rime mappings, a feasible approach is to work with within-word clusters that consist of three subsyllabic units, and with word-boundary clusters of two subsyllabic units. In principle, also other grain sizes could be used.

For example, if our polysyllabic word is composed of seven intrasyllabic units: u^1, u^2, u^3 ... u^7, and we want to work with clusters that are formed by three units, we thus construct the following OSLUs: ($u^1 + u^2$, i.e., the word's body), ($u^1 + u^2 + u^3$), ($u^2 + u^3 + u^4$), ... ($u^5 + u^6 + u^7$) and ($u^6 + u^7$, i.e., the word's rime). This procedure is demonstrated in Figures 3a and 3b.

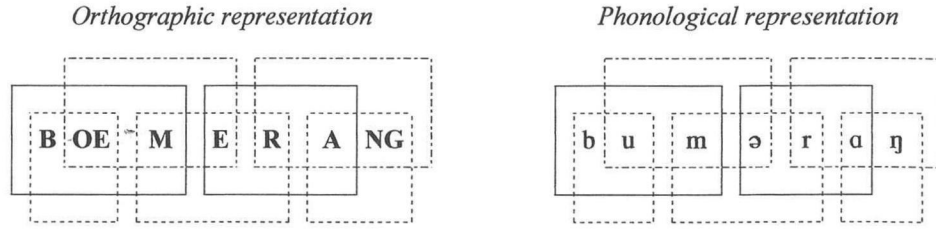

Figure 3a: *Overlapping SubLexical Units of a polysyllabic word*

---

and *WINTERSPORT*, whose different pronunciations of the cluster <rsp>, (/rsp/ and /rʃp/ respectively) reflect the different morpheme components of these words.

$O^1 + N^1$	<boe>	/bˈu/
$O^1 + N^1 + (C^1+O^2)$	<boem>	/bˈum/
$N^1 + (C^1+O^2) + N^2$	<oeme>	/ˈumə/
$(C^1+O^2) + N^2 + O^3$	<mer>	/mər/
$N^2 + O^3 + N^3$	<era>	/ərɑ/
$O^3 + N^3 + C^3$	<rang>	/rɑŋ/
$N^3 + C^3$	<ang>	/ɑŋ/

Figure 3b: *OLSU-structure of a polysyllabic word*

When we have segmented all the words in our corpus in this way, we check whether each of the clusters in our polysyllabic word is pronounced the same way it would be pronounced in other poly- or monosyllabic monomorphemic words, taking stress information, and, where applicable, syllable position,[10] tone and other prosodic features into account. In our example *BOEMERANG* (see Figure 4), we will have to compare the pronunciation of its seven OSLUs with the pronunciation of these OSLUs occurring in other Dutch monomorphemic words. Additional restrictions are that three OSLUs should contain a stressed vowel, marked by ˈ; and four OSLUs should contain an unstressed vowel, because we compare only stressed syllables with each other and unstressed with unstressed ones.

OSLUs of BOEMERANG		*Pronunciation of these OSLUs within other words*	
<boe>	/bˈu/	/bˈu/	*BOER, BOEF, BOEK, BOEL, BOETE,* . . .
<boem>	/bˈum/	/bˈum/	*BOEM, BOEMBOE-BOEMBOE,* . . .
<oeme>	/ˈumə/	/ˈumə/	*BOEMEL, NOEMER,* . . .
<mer>	/mər/	/mər/	*CAMERA, KAMERAAD,* . . .
<era>	/ərɑ/	/ərɑ/	*CHOLERA, PANGERANG,* . . .
<rang>	/rɑŋ/	/rɑŋ/	*BARANG, ORANG, PARANG,* . . .
<ang>	/ɑŋ/	/ɑŋ/	*MUSTANG, PISANG, SENANG,* . . .

Figure 4: *Comparing spelling-to-sound mappings for the OSLUs of BOEMERANG*

In our example, we see that there is no potential for inconsistency at any stage of the analysis. The pronunciation of *BOEMERANG* is entirely predictable, as shown when the spelling-to-sound mappings of its sublexical unit clusters are compared with those occurring in other words. One obvious advantage of this method is that it allows us to compare the consistent spelling and pronunciation of sublexical unit clusters without being restricted by the number of syllables our corpus material has.

If a cluster is unique (i.e., it only exists within one word), it does not disambiguate. Consequently, we proceed by comparing the mappings of the next OSLU. For example, suppose that we had not found another word containing the OSLU <era>. In that case, we would have proceeded to the next OSLU, looking for pronunciations of the pattern <rang>.

[10] This is necessary for those languages that have restrictions concerning the onset or coda of a word. E.g., the grapheme cluster <st> will be pronounced /ʃt/ in the onset of a German word like in *STEUER*, but the pronunciation is /st/ within a monomorphemic German word like *WESTE*.

After having compared the OSLUs in our word with the pronunciation or spelling in other words, we can assign a consistency score to our word. This score equals the number of occurrences of its 'friends' (i.e., words with analogous pronunciations) divided by the number of all occurrences for this pattern for each sublexical unit. If frequency information for the words is available, an additional consistency score can be computed: we can then divide the accumulated frequencies of its friends by the accumulated frequencies for this OSLU. The consistency of a sublexical cluster, thus, increases with the number of same mappings for the cluster or, analogously, with the frequency of the friends.

Considering that the OLSUs of *BOEMERANG* are pronounced in the same way in all other words in our corpus that contain these OSLUs, we can assign to our example *BOEMERANG* the complex consistency score '1111111', that is, a consistency measure for each OSLU, and a simple consistency score of '1', that is, the accumulated consistency score for each OSLU divided by the number of OSLUs. The first of these two consistency scores informs us at which part of the word an inconsistency is observed; the second provides us with an overall consistency measure for this word.

3. Application

In this section, we will illustrate with several examples how our method extends previous work on measuring consistency in several directions. First, the OSLU method can be applied to measure inconsistencies in sublexical unit clusters that are neglected by traditional analyses. Secondly, we will show the need for analyses of overlapping units for an analysis of the Russian language, as demonstrated by the existence of vowel graphemes that denote the palatalization of the preceding consonant. Finally, we will show that the predictive value of consistency measurements can be improved by incorporating data of polysyllabic words.

With our method, we can distinguish inconsistent onsets (like the spelling of /r/ in English, *RIGHT* or *WRONG*) from onsets that are ambiguous in isolation but that are consistent if the disambiguating context is taken into account. An example mentioned above is the pronunciation of <c> in English which becomes predictable when the nucleus of the syllable is taken into account.

Let us assume we want to determine the consistency of the sound-to-spelling mappings of the word *RIFT*. We know that the onset /r/ in isolation can be spelled in various ways, for example <r>, <rh> and <wr>. However, this ambiguous mapping might be disambiguated when the context is taken into account. Therefore, we first compare the mappings of the body /rɪ/. In our corpus, we find one single occurrence where the body is spelled as <rhy>, in *RHYTHM.* In 17 words in our sample, /rɪ/ is spelled as <ri> (*RICH*, *RING*, *RIBBON* . . .), whereas in 5 cases /rɪ/ is spelled as <wri> (*WRIGGLE*, *WRING*, *WRINKLE*, *WRIST* and *WRIT*). As a second step, we examine the mappings of the rimes.[11] Because the rime /ɪft/ is consistently spelled <ift>,

[11] In this example, we skipped a (trivial) step. Analogous to analyzing the mappings of the overlapping units of polysyllabic words, the next OSLU to be compared after comparing the bodies will have to be the word itself

we can conclude that the body cannot be written as <rhy> like in *RHYTHM*, as no orthographic rime <yft>, corresponding to the phonological rime /rft/, exists. Therefore, only two possibilities to spell /rɪft/ remain, namely <wrift> and <rift>. As a result, we can assign to *RIFT* the complex consistency score of '0.77/1' ('0.77' equals the number of analogous spellings, i.e. 17, divided by the number of occurrences of the orthographic body, i.e. 22; as the rime is unambiguous, it will be assigned the score '1'). Furthermore, we can assign *RIFT* the simple consistency score of '0.88' (both scores added and then divided by 2).

The procedure applied by traditional rime analyses assumes that we can assign each grapheme and phoneme to only one sublexical unit. One of the ensuing problems is the existence of graphemes that denote one phoneme and disambiguate another phoneme that does not belong to the same intrasyllabic unit. An example is the occurrence of some words with ambiguous onsets in British English, for instance, the phonological onset of *NEW*, /nju:/, is different from the onset in *NOON*, /nu:n/. Here, the pronunciation of the onsets is determined by the different orthographic rimes. While examples of this phenomenon are relatively rare in English, the phenomenon is a systematic characteristic of the Russian language where two sets of vowel graphemes exist: (<я>, <а>), (<ё>, <о>), (<ю>, <у>), (<е>, <э>), and (<и>, <ы>). The first member of each pair of graphemes denotes a vowel phoneme and the palatalization of the preceding consonant. The second member of the pair denotes the same vowel phoneme and the nonpalatalization of the preceding consonant. In contrast, pairs of two phonologically distinct consonants, one palatalized and the other nonpalatalized, are orthographically denoted by one single letter. The presence or absence of the palatalization is marked by the following vowels (or by one of two special letters, ъ, the 'hard sign', denoting the nonpalatalization, and ь, the 'soft sign', denoting the palatalization). Thus, the minimal pair люк /lʲuk/ vs. лук /luk/ is distinguished in its spoken form by different onsets /lʲ/ versus /l/. However, the corresponding written words do not deviate in the orthographic onset, but in the orthographic nucleus, while the phonological nuclei are identical. A rime analysis for Russian monosyllabic words that neglects the onsets and therefore does not capture the disambiguating function of the vowel would take Russian for a highly feedback inconsistent language because every phonological rime could generally correspond to two orthographic rimes.

A third way in which the presently proposed analysis can extend the range and accuracy of the traditional rime comparisons for monosyllabic words is to incorporate comparable OSLUs found in polysyllabic words into our analysis. In a truly consistent writing system, we expect a sublexical unit to be written and/or spelled in the same way depending on the stress structure but not depending on the number of syllables of the word, more specifically, on whether the word is monosyllabic or polysyllabic. Yet, as illustrated in the following example, such dependency on number of syllables appears to occur. The pronunciation of the orthographic rime <ice> in English, which is regarded consistent according to the traditional rime analysis that covers only monosyllabic words, is strictly speaking only consistent when taking monosyllabic words into account. In this case, we find only one pronunciation, namely /aɪs/. Comparing the pronunciation of all orthographic rimes <ice> occurring in a stressed sylla-

(onset, nucleus and coda). If, however, the word can be written or spelled in two ways, it is a homograph or homophone respectively, and its inconsistency is clearly shown.

ble,[12] we find a deviating pronunciation in a stressed syllable: in the word *POLICE* where the rime is pronounced /ˈi:s/. By expanding the lexical coverage to polysyllabic words, we can refine the consistency scores gathered by only considering monosyllabic words, and we can also measure the (in)consistency of some rimes that according to the more standard analyses were considered unique. For a lack of deviating pronunciations or spellings, the latter were automatically counted as consistent.

For example, by adding to our sample of 1149 monosyllabic German words the 1001 polysyllabic monomorphemic words with the stress on the last syllable and thus extending the corpus to 2150 rimes, in feedforward direction not 215 but 175 rimes remain unique; in feedback direction the number of unique rimes decreases from 130 to 112 rimes. In other words, the inclusion of polysyllabic words has the effect of decreasing the overestimation of consistency that can occur in small corpora. [13]

In this context it might be appropriate to deviate from the common procedure to categorize words as either consistent or inconsistent and instead establish some sort of threshold for the inconsistency status. As a consequence of the substantially greater number of sublexical units to be included into the analysis, there is a high chance that for many units a deviating pronunciation will be found. Strictly speaking, we would have to assign the feature 'inconsistent' to all OSLUs that could be pronounced or spelled in two ways. However, especially if the frequency of the deviating word is very low, or if the deviating word is clearly marked as a loan, it seems implausible that just a single divergent pronunciation or spelling can influence the consistency of all remaining analogous patterns. Setting up a threshold for consistency will, thus, prevent that the incidence of consistencies is underestimated.

Methods similar to the one proposed here are used in machine learning where algorithms analyze the correspondences between phonemes and graphemes in windows of varying sizes. As mentioned before, text-to-speech systems also use related methods to analyze the mappings of phonemes and letters in combination (i.e., diphones, triphones, bigrams, and trigrams). A major difference to the procedure proposed here is that in those approaches morphological and syllable structures are not taken into account. A language analyzed that way will be biased for consistency if it has a small inventory of graphemes formed by letter clusters (Van den Bosch et al. 1995). Additionally, the consistency results will come out higher for relatively isolating languages, as compared to languages like German where word formation processes like affixing and compounding are a major source of ambiguity of letter strings.[14] Determining the correct pronunciation or spelling for a word involves different subtasks, among others morphological segmentation, stress assignment, and graphemic parsing. While polygram- or polyphone-based methods like the above-mentioned can successfully predict the degree of overall (in)consistency of a language, research in the area of

---

[12] This is a necessary requirement, as the syllables of monosyllabic words are all carrying lexical stress.

[13] However, extending Ziegler's et al. (1997) corpus of 2.694 monomorphemic monosyllabic English words with all polysyllabic English words stressed on the last syllable resulted in only one single unique phonological rime that could be spelled in a deviating way in a polysyllabic word.

[14] Examples are the letter sequences <ng> and <sch> whose pronunciations are disambiguated by morphological segmentation, as in *ANGEBEN*, /ng/ vs. *ANGELN*, /ŋ/ and *HÄUSCHEN*, /sç/ vs. *TÄUSCHEN*, /ʃ/. In contrast, morphological decomposition is much less an issue for another inflecting and compounding language, Dutch, as morpheme boundaries are less restrictive to assimilation processes.

psycholinguistics often aims at disentangling the different components mentioned above, and at analyzing them in isolation. Therefore, our method measures the consistency of only one component, namely the pronunciation of monomorphemic words.

4. Conclusion

The OSLU-approach presents an alternative to the standard methods to measure the degree of spelling-sound consistency. Instead of focusing on the rimes of monosyllabic monomorphemic words, our method determines the degree of consistency of all sublexical units of both poly- and monosyllabic monomorphemic words. The procedure is language-independent, relatively simple, variable in terms of unit size, and introduces, as far as we know, a new method to measure the overall spelling-sound consistency. Expanding the analysis to polysyllabic words produces a much more complete and less biased picture of the orthographic depth of a language. The appendix presents the results of sample analyses of the Dutch and German Celex corpora that illustrate how the method described in the article works.

In addition to merely computing the percentages of (in)consistent spellings or pronunciations in a particular language corpus, it is desirable that further research on phonological consistency also involves the study of the nature of the inconsistencies that are observed. For example, due to the regular phonological process of final devoicing, word-final obstruent consonants are devoiced in German. Consequently, many rimes are feedback inconsistent, for example the rimes of *WALD*, /valt/ vs. *KALT*, /kalt/. A native speaker of German, however, can produce the correct spelling on the basis of the knowledge that the plural form of *WALD* is *WÄLDER*, /vˈɛldər/. Other German words, for example *HERBST*, cannot be disambiguated by the corresponding paradigm, and, therefore, pose a 'real' problem for spelling.

It may also be informative to measure the degree of inconsistency: Feature-based distance metrics between phonemes could determine the degree of the difference between two different pronunciations for the same pattern (Cucchiarini 1993). While in German, the inconsistent rime pronunciations of the words *DACH*, /dɑx/ and *SCHMACH*, /ʃmaːx/ only differ in vowel length (and, consequently, in tenseness), the different pronunciations of English words ending with the orthographic rime <ough> show much larger differences in their pronunciations (e.g., *BOUGH*, *COUGH*, *DOUGH* and *ROUGH)*.

Furthermore, it may be interesting to explore the possibility to predict feedforward and feedback inconsistencies without actually performing a detailed corpus analysis. Because the pronunciation and spelling of loan words constitutes one of the major sources of inconsistencies, it might be worthwhile to test their integration into the native system. A small sample of concepts, typically expressed by international words, could be analyzed cross-linguistically in order to see to what degree these words are adapted to the graphemic, phonemic and morphological structure of the borrowing languages. This approach is, for instance, adopted by Meisenburg (1992).

The choice of a specific method to measure the phonological consistency of a language is inherently independent of assumptions about the actual processing of the language. However,

after the presently proposed procedure has been applied on a suitably large corpus, analogously to the analyses performed by Ziegler et al. (1996, 1997) and Martensen et al. (2000), the effect of the inconsistencies as determined by our method on spelling and reading performance can subsequently be tested. This way, the psychological validity of the presented (in)consistency measurement could be verified in empirical studies.

References

Baayen, R.H., R. Piepenbrock & L. Gulikers (1995): *The CELEX Lexical Database.* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baron, J., R. Treiman, J.J. Freyd & P. Kellman (1980): Spelling and Reading by Rules. In: U. Frith (ed.): *Cognitive Processes in Spelling.* London: Academic Press, 159-194.

Berndt, R.S., J.A. Reggia & C.C. Mitchum (1987): Empirically Derived Probabilities for Grapheme-to-Phoneme Correspondences in English. *Behavior Research Methods, Instruments, & Computers* 19: 1-9.

Borgwaldt, S.R., & A.M.B. De Groot (in preparation): *Spelling-Sound Relations in English, Dutch and German.* Ms. University of Amsterdam.

Brown, G.D.A. & N.C. Ellis (1994): Issues in Spelling Research: An Overview. In: G.D.A. Brown & N.C. Ellis (eds.): *Handbook of Spelling: Theory, Process and Intervention.* New York: John Wiley & Sons, 3-25.

Cucchiarini, C. (1993): *Phonetic Transcription: A Methodological and Empirical Study.* Doctoral thesis, University of Nijmegen.

Dewey, G. (1971): *English Spelling: Roadblock to Reading.* New York: Teachers College Press.

Geilfuß-Wolfgang, J. (this volume): *Optimal Hyphenation.*

Glushko, R.J. (1979): The Organization and Activation of Orthographic Knowledge in Reading Aloud. *Journal of Experimental Psychology: Human Perception and Performance* 5: 674-691.

Iverson, G.K. & D.W. Wheeler (1989): Phonological Categories and Constituents. In: R. Corrigan (ed.): *Linguistic Categorization.* Amsterdam: Benjamins, 93-114.

Karlsson, F. (1983): *Suomen kielen äänne– ja muotorakenne.* Juva: WSOY.

Kessler, B. & R. Treiman (2001): Relations Between Sounds and Letters in English Monosyllables. *Journal of Memory and Language* 44: 592-617.

Marchand, Y. & R.I. Damper (2000): A Multi-Strategy Approach to Improving Pronunciation by Analogy. *Computational Linguistics* 26: 195-219.

Martensen, H., E. Maris & T. Dijkstra (2000): When does Inconsistency Hurt? On the Relation between Consistency Effects and Reliability. *Memory & Cognition* 28: 648-656.

Meisenburg, T. (1992): Graphische und phonische Integration von Fremdwörtern am Beispiel des Spanischen. *Zeitschrift für Sprachwissenschaft* 11: 47-67.

Neef, M. (this volume): *The Reader's View: Sharpening in German.*

Nunn, A.M. (1998): *Dutch Orthography. A Systematic Investigation of the Spelling of Dutch Words.* The Hague: Holland Academic Graphics.

Sproat, R. (2000): *A Computational Theory of Writing Systems.* Stanford, CA: Cambridge University Press.

– (this volume): The Consistency of the Orthographically Relevant Level in Dutch.

Stone, G.O., M. Vanhoy & G.C. Van Orden (1997): Perception is a Two-Way Street: Feedforward and Feedback Phonology in Visual Word Recognition. *Journal of Memory and Language* 36: 337-359.

Treiman, R., J. Mullennix, R. Bijeljac-Babic & E.D. Richmond-Welty (1995): The Special Role of Rimes in the Description, Use, and Acquisition of English Orthography. *Journal of Experimental Psychology: General* 124: 107-136.

Van den Bosch, A., A. Content, W. Daelemans & B. de Gelder (1995): Measuring the Complexity of Writing Systems. *Journal of Quantitative Linguistics* 1: 178-188.

Venezky, R.L. (1970): *The Structure of English Orthography.* The Hague: Mouton.

Véronis, J. (1986): Etude quantitative sur le système graphique et phonologique du français. *Cahiers de Psychologie Cognitive* 6: 501-531.

Yvon, F. (1997): Paradigmatic Cascades: A Linguistically Sound Model of Pronunciation by Analogy. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, Madrid, July 1997.

Ziegler, J.C., A.M. Jacobs & G.O. Stone (1996): Statistical Analysis of the Bidirectional Inconsistency of Spelling and Sound in French. *Behavior Research Methods, Instruments, & Computers* 28: 504-515.

Ziegler, J.C., G.O. Stone & A.M. Jacobs (1997): What's the Pronunciation for *-ough* and the Spelling for /u/? A Database for Computing Feedforward and Feedback Inconsistency in English. *Behavior Research Methods, Instruments, & Computers* 29: 600-618.

Appendix

Sample analyses on the first 50 words of a corpus of all monomorphemic Dutch and German words in the Celex corpus.

The first column contains the orthographic/phonological forms of the words. The seconds column represents the global consistency scores, 'ff' in feedforward direction and 'fb' in feedback direction. The third column contains the orthographic/phonological OSLUs, with different pronunciations/spellings for these OSLUs indicated, together with the number of occurrences, in brackets.

Dutch subset (only constraint: final syllables)

Word/ Pronunciation	Score		OSLUs With Correspondences
aai	ff	1	aai
'a:j	fb	1	'a:j
aaien	ff	1	aai, aaie, ien
'a:jə	fb	1	'a:j, 'a:jə, jə
aak	ff	1	aak
'a:k	fb	1	'a:k
aal	ff	1	aal
'a:l	fb	1	'a:l
aan	ff	1	aan
'a:n	fb	0.976	'a:n [<aan>: 41; <an>: 1]
aap	ff	1	aap
'a:p	fb	1	'a:p
aar	ff	1	aar
'a:r	fb	0.678	'a:r [<aar>: 38; <oir>: 12, <ard>: 4, <oire>: 2]
aard	ff	1	aard
'a:rt	fb	0.642	'a:rt [<aard>: 9; <aart>: 5]
aarde	ff	1	aard, aarde, rde
'a:rdə	fb	0.737	'a:rd, 'a:rdə [<aarde>: 2; <aarden>: 1], rdə [<rde>: 6; <rden>: 5]
aarden	ff	1	aard, aarde, rden
'a:rdə	fb	0.595	'a:rd, 'a:rdə [<aarde>: 2; <aarden>: 1], rdə [<rde>: 6; <rden>: 5]
aardig	ff	1	aard, aardi, rdig, ig
'a:rdəx	fb	1	'a:rd, 'a:rdə, rdəx, əx
aars	ff	1	aars
'a:rs	fb	1	'a:rs
aarzelen	ff	1	aarz, aarze, rzel, ele, len
'a:rzələ	fb	1	'a:rz, 'a:rzə, rzəl, ələ, lə

aas	ff	1	aas
'a:s	fb	1	'a:s
abattoir	ff	1	ab, aba, batt, attoi, ttoir, oir
a:bɑtw'a:r	fb	0.934	a:b, a:bɑ, bɑtw, ɑtw'a:, tw'a:r [<toir>: 6; <toirē>: 2], 'a:r [<oir>: 12; <toire>: 2]
abbé	ff	1	abb, abbé, bbé
ɑb'e:	fb	0.555	ɑb [<ab>: 5; <abb>: 1], ɑb'e:, b'e: [<bbé>: 1; <bee>: 1]
abces	ff	1	abc, abce, bces, es
ɑps'ɛs	fb	0.958	ɑps [<abc>: 1; <abs>: 5], ɑps'ɛ, ps'ɛs, 'ɛs
abdij	ff	1	abd, abdij, bdij
ɑbd'ɛi	fb	1	ɑbd, ɑbdɛi, bdɛi
abel	ff	0.952	ab[/'a:b/: 15; /'e:b/: 2], abe [/'a:bə/: 13; /'e:bə/: 1], bel, el
'a:bəl	fb	1	'a:b, 'a:bə, bəl, əl
aberratie	ff	1	ab, abe, berr, erra, rrat, atie, tie
ɑbər'a:tsi:	fb	1	ɑb, ɑbə, bər, ər'a:, r'a:ts, 'a:tsi:, tsi:
abituriënt	ff	1	ab, abi, bit, itu, tur, uri, rië, iënt, ënt
ɑbi:ty:ri:j'ɛnt	fb	1	ɑb, ɑbi:, bi:t, i:ty:, ty:r, y:ri:, i:j'ɛ, j'ɛnt, 'ɛnt,
abonnee	ff	1	ab, abo, bonn, onnee, nnee
ɑbɔn'e:	fb	1	ɑb, ɑbɔ, bɔn, ɔn'e:, n'e:
abonneren	ff	1	ab, abo, bonn, onne, nner, ere, ren
ɑbɔn'e:rə	fb	1	ɑb, ɑbɔ, bɔn, ɔn'e:, n'e:r, 'e:rə, rə
aborigines	ff	1	ab, abo, bor, ori, rig, igi, gin, ine, nes, es
a:bo:r'i:dʒi:nɛs	fb	1	a:b, a:bo:, bo:r, o:r'i:, r'i:dʒ, 'i:dʒi:, dʒi:n, i:nɛ, nɛs, ɛs
abortus	ff	1	ab, abo , bort, ortu, rtus, us
a:b'ɔrtʏs	fb	1	a:b, a:b'ɔ, b'ɔrt, 'ɔrtʏ, rtʏs, ʏs
abrikoos	ff	1	abr, abri, brik, ikoo, koos, oos
ɑbri:k'o:s	fb	0.916	ɑbr, ɑbri:, bri:k, i:k'o, [<ikoo>: 1, <icoo>: 1], k'o:s, 'o:s
abrupt	ff	1	abr, abru, brupt, upt
ɑbr'ʉpt	fb	1	ɑbr, ɑbr'ʏ, br'ʏpt, 'ʏpt
absolutisme	ff	1	abs, abso, bsol, olu, lut, uti, tism, isme, sme
ɑpso:ly:t'ɪsmə	fb	1	ɑps, ɑpso:, pso:l, o:ly:, ly:t, y:t'ɪ, t'ɪsm, 'ɪsmə, smə
absoluut	ff	1	abs, abso, bsol, oluu, luut, uut
ɑpso:l'y:t	fb	1	ɑps, ɑpso:, pso:l, o:l'y:, l'y:t, 'y:t
absorberen	ff	1	abs, abso, bsorb, orbe, rber, ere, ren
ɑpsɔrb'e:rə	fb	0.976	ɑps [<abc>: 1; <abs>: 5], ɑpsɔ, psɔrb, ɔrb'e:, rb'e:, 'e:rə, rə
absorptie	ff	1	abs, abso, bsorpt, orptie, rptie
ɑps'ɔrpsi:	fb	0.966	ɑps [<abc>: 1; <abs>: 5], ɑps'ɔ, ps'ɔrps, 'ɔrpsi:, rpsi:
abstract	ff	1	abstr, abstra, bstract, act
ɑpstr'ɑkt	fb	1	ɑpstr, ɑpstr'ɑ, pstr'ɑkt, 'ɑkt
absurd	ff	1	abs, absu, bsurd, urd
ɑps'ʉrt	fb	0.874	ɑps [<abc>: 1; <abs>: 5], ɑps'ʉ, ps'ʉrt, 'ʉrt [<urd>: 1; <irt>: 2]

abt	ff	1	abt
'ɑpt	fb	1	'ɑpt
abuis	ff	1	ab, abui, buis, uis
a:b'œys	fb	1	a:b, a:b'œy, b'œys, 'œys
acacia	ff	1	ac, aca, cac, aci, cia, ia, a
a:k'a:si:ja:	fb		a:k, a:k'a:, k'a:s, 'a:si:, si:ja:, i:ja:, ja:
academicus	ff	1	ac, aca, cad, ade, dem, emi, mic, icu, cus, us
a:ka:d'e:mi:kʉs	fb		a:k, a:ka:, ka:d, a:d'e:, d'e:m, 'e:mi:, mi:k, i:kʏ, kʏs, ʏs
accent	ff	1	acc, acce, ccent, ent
ɑks'ɛnt	fb	0.987 5	ɑks, ɑks'ɛ, ks'ɛnt, 'ɛnt [<ent>: 38; <end>: 2]
accident	ff	1	acc, acci, ccid, ide, dent, ent
ɑksi:d'ɛnt	fb	0.797	ɑks [<acc>: 3; <ax>: 1], ɑksi: [<acci>: 3; <axi>: 1], ksi:d [<ccid>: 1; <xid>: 1], i:d'ɛ, d'ɛnt [<dent>: 5; <dend>: 1], 'ɛnt [<ent>: 38; <end>: 2]
accijns	ff	1	acc, accij, ccijns, ijns
ɑks'ɛins	fb	0.937	ɑks [<acc>: 3; <ax>: 1], ɑks'ɛi, ks'ɛins, 'ɛins
accordeon	ff	1	acc, acco, ccord, orde, rdeo, eon, on
ɑkɔrde:j'ɔn	fb	1	ɑk, ɑkɔ, kɔrd, ɔrde:, rde:j'ɔ, e:j'ɔn, j'ɔn
accountant	ff	1	acc, accou, ccount, ounta, ntant, ant
ɑk'ɑuntənt	fb	1	ɑk, ɑk'ɑu, k'ɑunt, 'ɑuntə, ntənt, ənt
accu	ff	1	acc, accu, ccu
'ɑky:	fb	1	'ɑk, 'ɑky:, ky:
acculturatie	ff	1	acc, accu, ccult, ultu, ltur, ura, rat, atie, tie
ɑkʉlty:r'a:tsi:	fb	0.823	ɑk [<ac>: 2; <acc>: 5], ɑkʉ, [<acu>: 1; <accu>: 1], kʉlt, [<ccult>: 1; <cult>: 4], ʉlty:, lty:r, y:r'a:, r'a:ts, 'a:tsi:, tsi:
accumuleren	ff	1	acc, accu, ccum, umu, mul, ule, ere, ren
ɑky:my:l'e:rə	fb	0.780	ɑk [<ac>: 2; <acc>: 5], ɑky: [<acu>: 1; <accu>: 1], ky:m [<acu>: 1; <accu>: 1], y:my:, my:l, y:l'e:, 'e:rə, rə
accuraat	ff	1	acc, accu, ccur, uraa, raat, aat
ɑky:r'a:t	fb	0.669	ɑk [<ac>: 2; <acc>: 5], ɑky: [<acu>: 1; <accu>: 1], ky:r [<cur>: 7; <ccur>: 1], y:r'a:, r'a:t [<raat>: 9; <aad>: 2], 'a:t [<aat>: 66; <aad>: 11],
acht	ff	1	acht
'ɑxt	fb	0.941	'ɑxt [<axt>: 16; <agd>: 1]
achten	ff	1	acht, achte, chten
'ɑxtə	fb	1	'ɑxt, 'ɑxtə, xtə
acne	ff	1	acn, acne, cne
'ɑknə	fb	1	'ɑkn, 'ɑknə, knə
acrobaat	ff	1	acr, acro, crob, obaa, baat, aat
ɑkro:b'a:t	fb	1	ɑkr, ɑkro:, kro:b, o:b'a:, b'a:t, 'a:t

German subset

Word/ Pronunciation	Score		OSLUs with Correspondances
aal	ff	1	aal
'a:l	fb	0.029	'a:l [<aal>: 2; <ahl>: 7; <al>: 58]
aas	ff	1	aas
'a:s	fb	0.125	'a:s [<aas>: 1; <aß>: 3; <as>: 4]
abend	ff	1	ab, abe, bend, end
'a:bənt	fb	0.66	'a:b, 'a:bə, bənt, ənt [<end>: 6; <ent>: 3]
abitur	ff	1	ab, abi, bit, itu, tur, ur
abi:t'u:r	fb	0.981	ab, abi:, bi:t, i:t'u, t'u:r [<tur>: 17; <tour>: 1], 'u:r [<ur>: 33; <our>: 2]
abonnement	ff	0.875	ab, abo, bonn, onne, nnem, eme, ment [/m'ɛnt/: 21; /m'ɑ̃:/: 7]
abɔnəm'ɑ̃:	fb	1	ab, abɔ, bɔn, ɔnə, nəm, əm'ɑ̃:, m'ɑ̃:
abonnent	ff	0.855	ab, abo, bonn, onne, nnent, ent [/'ɛnt/: 46; /ɑ̃:/: 7]
abɔn'ɛnt	fb	1	ab, abɔ, bɔn, ɔn'ɛ, n'ɛnt, 'ɛnt
abonnieren	ff	1	ab, abo, bonn, onnie, nnier, iere, ren, en
abɔn'i:rən	fb	1	ab, abɔ, bɔn, ɔn'i:, n'i:r, 'i:rə, rən, ən
absolut	ff	1	abs, abso, bsol, olu, lut, ut
apzo:l'u:t	fb	1	apz, apzo:, pzo:l, o:l'u:, l'u:t, 'u:t
absorption	ff	1	abs, abso, bsorpt, orpti, rptio, ion, on
apzɔrptsi:'o:n	fb	1	apz, apzɔ, pzɔrpt, ɔrptsi:, rptsi:'o:, i:'o:n, 'o:n
abstrahieren	ff	1	abstr, abstra, bstrah, ahie, hier, iere, ren, en
apstrah'i:rən	fb	0.998	apstr, apstra, pstrah, ah'i:, h'i:r, 'i:rə [<iere>: 90; <ire>: 1], rən, ən,
abstrakt	ff	1	abstr, abstra, bstrakt, akt
apstr'akt	fb	0.964	apstr, apstr'a, pstr'akt, 'akt [<akt>: 12; <ackt>: 1, <agd>: 1]
abstrus	ff	0.928	abstr, abstru, bstrus, us [/'ʊs/: 2; /'u:s/: 5]
apstr'u:s	fb	0.875	apstr, apstr'u:, pstr'u:s, 'u:s [<ues>: 1; <us>: 4, <uß>: 3]
absurd	ff	1	abs, absu, bsurd, urd
apz'ʊrt	fb	0.875	apz, apzʊ, pzʊrt, ʊrt [<urt>: 1; <urd>: 1]
abt	ff	1	abt
'apt	fb	1	apt
achse	ff	1	achs, achse, chse
'aksə	fb	0.666	'aks [<achs>: 7; <ax>: 5], 'aksə [<achse>: 4; <axe>: 2], ksə [<chse>: 9; <xe>: 3]
achsel	ff	1	achs, achse, chsel, el
'aksəl	fb	1	'aks, 'aksə, ksəl, ksəl, əl
acht	ff	1	acht
'axt	fb	1	'axt

achten	ff	1	acht, achte, chten, en
'axtən	fb	1	'axt, 'axtə, xtən, ən
ächzen	ff	1	ächz, ächze, chzen, en
'ɛxtsən	fb	0.916	'ɛxts, 'ɛxtsə [<echze>: 1; <ächze>: 2], xtsən, ən
adäquat	ff	0.996	ad, adä, däqu, äqua, quat, at [/a:t/: 44; /a:/: 1]
adɛkv'a:t	fb	1	ad, adɛ, dɛkv, dɛkv'a:, ɛkv'a:, kv'a:t, a:t
adel	ff	1	ad, ade, del, el
'a:dəl	fb	1	'a:d, 'a:də, dəl, dəl, əl
adeln	ff	1	ad, ade, deln, eln
'a:d'əln	fb	1	'a:d, 'a:də, dəln, dəln
ader	ff	1	ad, ade, der, er
'a:dər	fb	1	'a:d, 'a:də, dər, ər
adjustieren	ff	1	adj, adju, djust, ustie, stier, iere, ren, en
atjʊst'i:rən	fb	0.998	atj, atjʊ, tjʊst, ʊst'i:, st'i:r, 'i:rə [<iere>: 90; <ire>: 1], rən, ən
adjutant	ff	1	adj, adju, djut, uta, tant, ant
atju:t'ant	fb	1	atj, atju:, tju:t, u:t'a, t'ant, 'ant
adler	ff	1	adl, adle, dler, er
'a:dlər	fb	1	'a:dl, 'a:dlə, dlər, ər
adlige	ff	1	adl, adli, dlig, ige, ge
'a:dlɪgə	fb	1	'a:dl, 'a:dlɪ, dlɪg, ɪgə, gə
admiral	ff	1	adm, admi, dmir, ira, ral, al
atmi:r'a:l	fb	1	atm, atmi:, tmi:r, i:r'a:, r'a:l, 'a:l
adrett	ff	1	adr, adre, drett, ett
adr'ɛt	fb	0.991	adr, adr'ɛ, dr'ɛt, 'ɛt [<ett>: 29; <et>: 1]
advent	ff	0.966	adv, adve, dvent, ent [/'ɛnt/: 46; /'ã:/: 7]
atv'ɛnt	fb	0.984	atv, atv'ɛ, tv'ɛnt, v'ɛnt, 'ɛnt [<ent>: 45; <and>: 1; <end>: 2]
adverb	ff	1	adv, adve, dverb, erb
atv'ɛrp	fb	1	atv, atv'ɛ, tv'ɛrp, 'ɛrp
advokat	ff	1	adv, advo, dvok, oka, kat, at
atvo:k'a:t	fb	1	atv, atvo:, tvo:k, o:k'a:, k'a:t, 'a:t
affäre	ff	1	aff, affä, ffär, äre, re
af'ɛ:rə	fb	1	af, af'ɛ:, f'ɛ:r, 'ɛ:rə, rə
affe	ff	1	aff, affe, ffe
'afə	fb	1	'af, 'afə, fə
affekt	ff	1	aff, affe, ffekt, ekt
af'ɛkt	fb	1	af, af'ɛ, f'ɛkt, 'ɛkt
äffen	ff	1	äff, äffe, ffen
'ɛfən	fb	0.633	'ɛf [<äff>: 2; <eff>: 3], 'ɛfə [<äffe>: 2; <effe>: 2], fən
affront	ff	0.8125	affr, affro, ffront, ont [<õ:>: 1; <ɔnt>: 3]
afr'õ:	fb	1	afr, afr'õ:, fr'õ:

agave	ff	1	ag, aga, gav, ave, ve
ag'a:və	fb	1	ag, ag'a:, g'a:v, 'a:və, və
agens	ff	1	ag, age, gens, ens
'a:gəns	fb	1	'a:g, 'a:gə, gəns, əns
agent	ff	1	ag, age, gent, ent
ag'ɛnt	fb	1	ag, ag'ɛ, g'ɛnt, 'ɛnt
aggression	ff	1	aggr, aggre, ggress, essi, ssio, ion, on
agrɛsi:'o:n	fb	0.928	agr [<aggr>: 1; <agr>: 1], agrɛ, grɛs, ɛsi:, si:'o:, i:'o:n, 'o:n
ägide	ff	1	äg, ägi, gid, ide, de
ɛ:g'i:də	fb	0.971	ɛ:g, ɛ:g'i:, g'i:d, 'i:də [<iede>: 1; <ide>: 6], də
agil	ff	0.998	ag, agi, gil, il [/'i:l/: 21; /ɪl/: 1],
ag'i:l	fb	0.875	ag, ag'i: [<agie>: 1; <agi>: 1], g'i:l, 'i:l [<il>: 21; <iel>: 3]
agronom	ff	1	agr, agro, gron, ono, nom, om
agro:n'o:m	fb	1	agr, agro:, gro:n, o:n'o:, n'o:m, 'o:m
ahle	ff	1	ahl, ahle, le
'a:lə	fb	0.444	'a:l [<ahl>: 7; <al>: 58], 'a:lə [<ahle>: 5; <ale>: 17], lə
ahmen	ff	1	ahm, ahme, men, en
'a:mən	fb	0.633	'a:m [<am>: 20; <ahm>: 5], 'a:mə [<ame>: 6; <ahme>: 3], mən, ən
ahn	ff	1	ahn
'a:n	fb	0.139	'a:n [<an>: 37; <ahn>: 6]
ahnden	ff	1	ahnd, ahnde, nden, en
'a:ndən	fb	1	'a:nd, 'a:ndə, ndən, ən
ähneln	ff	1	ähn, ähne, neln, eln
'ɛ:nəln	fb	0.666	'ɛ:n [<än>: 1; <ähn>: 1], 'ɛ:nə [<äne>: 5; <ähne>: 4], nəln, əln
ahnen	ff	1	ahn, ahne, nen, en
'a:nən	fb	0.638	'a:n [<an>: 37; <ahn>: 6], 'a:nə [<ane>: 7; <ahne>: 5], nən, ən