Spelling space: A computational testbed for phonological and morphological changes in Dutch spelling

Antal van den Bosch
ILK / Dept. of Language and Information Science
Faculty of Arts, Tilburg University
P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands
Antal.vdnBosch@uvt.nl

Abstract

The Dutch spelling system, like other European spelling systems, represents a certain balance between preserving the spelling of morphemes (the morphological principle) and obeying letter-to-sound regularities (the phonological principle). We present experimental results with artificial learners that show a competition effect between the two principles: adhering more to one principle leads to more violations of the other. The artificial learners, memory-based learning algorithms, are trained (1) to convert written words to their phonemic counterparts and (2) to analyze written words on their morphological composition, based on data extracted from the CELEX lexical database. As an exception to the competition effect we show that introducing the schwa as a letter in the spelling system causes both morphology and phonology to be learnt better by the artificial learners. In general we argue that artificial learning studies are a tool in obtaining objective measurements on a spelling system that may be of help in spelling reform processes.

1. Introduction

The Dutch spelling system has emerged from a long evolution. Increasingly by explicit design, but also before governmental initiatives began to influence it, Dutch spelling has been evolving to a state in which on the one hand the spelling of morphemes is preserved in the way words are written, while on the other hand written forms tend to obey letter-to-sound regularities – mostly not one-to-one regularities, but governed by simple rules with limited context-sensitivity. All European alphabetic writing systems exhibit some form of balance between preserving the spelling of morphemes, and obeying letter-to-sound regularities (Raible, 1991).

To a certain extent, the goals of preserving morphology and following phonology conflict; in many wordforms, preserving the morphological structure in the surface wordform implies a violation of letter-to-sound regularities, and vice versa. The situation can be likened to a tug-of-war of approximately equally strong parties, where the spelling is the rope, and morphology and phonology are pulling it, as visualized in Figure 1.

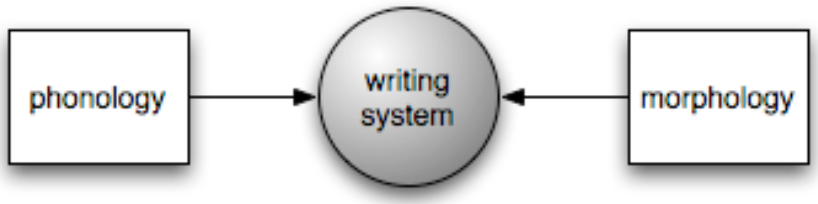

Figure 1. Visualization of the balanced "tug-of-war" between phonology and morphology; by exerting the same force, the spelling system remains in the middle. Should either side become stronger, spelling would go more in the direction of the stronger party, and further away from the other party.

In this article we present experimental results that show that when spelling is made more phonological, the morphological information in the surface wordforms becomes harder to access, and vice versa, indicating that indeed there is a tense balance between the two. We do this by training artificial learners from machine learning, in particular memory-based learners (Daelemans and Van den Bosch, 2005) to transliterate written words to phonemic transcriptions, and to perform morphological analyses on the same written words. When the correspondence between written and phonemic forms becomes more regular, artificial learners are more successful in learning to transliterate – measurable by testing the learner after training on unseen words. The same goes for the correspondence between written words and their analyses.

The article is structured as follows. In this introductory section we provide a brief overview of Dutch spelling, and provide examples of the tension between phonology and morphology in orthographical Dutch words. In Section 2 we introduce the artificial learner used throughout this study, and with this learner we determine the "current" situation of Dutch spelling in order to establish baseline results to compare later results with. In Section 3 we experimentally change Dutch spelling to being fully regular in the mapping between letters and phonemes, and measure the learnability of both letter-phoneme transliteration and morphological analysis after this change. In Section 4 we describe the converse experiment in which words are artificially changed to retain the surface spelling of all morphemes. Section 5 subsequently presents an experiment in which the schwa phoneme is introduced as a letter, causing both morphology and phonology to be learnt better by the artificial learners. In Section 6 we summarize our findings, voice our conclusions and mention points for further research.

1.1. The tension between phonology and morphology in current Dutch spelling

The current Dutch spelling system is, as most of its neighboring West-European systems, a product of many different evolutions. As Raible (1991)

shows, European alphabetic writing systems such as that of Dutch have always tended to drift away from purely phonological spellings towards systems that incorporate elements that help the reader in understanding writing, such as word space, the distinction between small and capital letters, the introduction of punctuation, and the concept of stem constancy. We refer to the latter as the *morphological principle*: words should retain a constant spelling of stems. In contrast, the *phonological principle* dictates that the relations between the spelling of words and their pronunciations be as direct as possible. These principles tend to clash. We identify two classes of principle clashes in current Dutch spelling in which more than one principle applies, but one wins:

1. The phonological principle applies, but the morphological principle would predict a different spelling.

2. The morphological principle applies, but the phonological principle would predict a different spelling.

We briefly describe each class by providing examples.

1.2 Phonology wins over morphology

Two rather frequent clashes in which both the phonological principle and the morphological principle apply, and the phonological principle wins, occur with particular cases of voicing and gemination triggered by morphological inflections, and spelling of long vowels triggered by syllable structure.

The <-en> inflection is implicated both in many cases of plural formation of nouns, and in plural person, present tense verb inflections, as well as in infinitive verb inflection. It is a very common morpheme, pronounced usually as [ə]. It tends to voice [s] and [f] codas following a long vowel or diphthong nucleus in the last syllable of the word it attaches to, to [z] and [v], respectively. For example, it changes the pronunciation of the <s> in <muis> `mouse' to [z] in <muizen> `mice'. Reflecting the pronunciation, the spelling contains a <z>, whereas the morphological principle would have wanted the spelling to be <muisen>, retaining the spelling of the singular-noun stem <muis>. Gemination (doubling) of consonant letters often occurs in complimentary cases when the <-en> inflection attaches to a word ending in a consonant after a short vowel. The singular form <bus> (bus) doubles the <s> in the plural form <bussen> `buses'. Both phenomena also occur with the frequent adjectival <-e> inflection on prenominal adjectives (<braaf> `good', versus prenominal <brave>), and with the verbal <-end> and <-ende> inflections for present participles (<graas> `graze', versus <grazend> `grazing'; <bak> `bake', versus <bakkend> `baking'). The two phenomena are quite pervasive.

A third case of phonology winning over morphology is the non-constant spelling of long vowels ([a], [e], [i], [o], [u]). They are either spelled with two letters or with one, depending on syllable structure. In a closed syllable ending with one or more consonants, a long vowel is spelled with two letters (e.g. <oo> in <loop> `walk', singular), but in an open syllable ending in the long vowel, only one letter is used, such as in <lopen> (plural), in which the <p> has become the onset of the final syllable due to the plural inflection.

1.3. Morphology wins over phonology

Introduced as the morphological principle, the spelling of certain stems are retained in all of their paradigmatic wordforms, regardless of the inflections that may attach to them. This leads to odd consonantal clusters typical for Dutch, such as <dt>, when a <t> verbal inflection marking the second or third person singular is attached to a verb stem ending in <d> (e.g., leading to <vindt> `finds', third or second person singular) that are not pronounced as such (<vindt> is pronounced as [vɪnt]).

Cases in which the spelling of morphemes is retained, and in which the aforementioned voicing of <f> into <v> and <s> into <z> is blocked, is in compounds. For example, the compound noun <asbak> `ash tray' may be pronounced [ɑzbɑk], but the spelling holds on to <s>, retaining the spelling of the stem <as> (ash).

2. Spelling space: Current Dutch

In this section we analyse the learnability of phonemic transliteration and morphological analysis of "current" Dutch – we specify which snapshot and sample of Dutch we use in our experiments. We then provide technical background information regarding the artificial learners and the two processing tasks. We then present the results in a two-dimensional space we call "spelling space".

2.1. A snapshot and sample of Dutch

For our experiments we have used the Dutch CELEX lexical database (Baayen et al., 1993). In time, this puts us before the relatively minor changes incurred by the "spelling-Geerts" of 1995, and an upcoming refinement of this spelling in 2006. CELEX offers, among other information fields, phonemic transcriptions (in the SAMPA phonemic alphabet) and morphological analyses of Dutch words – among which a considerable amount of automatically generated, hand-checked inflections. For 336,698 words, both a phonemic transcription and a morphological analysis are available; we based our experiments on this set of words. Within these 336,698 words, 293,825

unique phonemizations occur – many words have multiple morphological analyses, but a single pronunciation. The other way around (in which words with multiple pronunciations share the same morphology) does not occur, except in stress patterns, which we disregard in the present study.

2.2. An artificial learner: Memory-based learning

As the artificial learner in our experiments we use memory-based learning (Daelemans and Van den Bosch, 2005), also known as the k-nearest neighbor classifier (Cover & Hart, 1967) or instance-based learning (Aha et al., 1991). Memory-based learning has two distinguishable parts: a learner and a processing module (a classifier). Learning consists of storing individual examples of a processing task in memory. Each example represents a mapping from input (e.g. letters) to output (e.g. a phoneme mapping to one of the input letters). Classification involves the matching of a new, unseen example to all examples in memory, and extrapolating the majority class of the nearest-neighbor examples to the new example. The k parameter determines the number of nearest neighbors used in classification; we set k = 1. Nearest neighbors are ranked according to the distance function between two instances X and Y, $\Delta(X,Y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i)$, where *n* is the number of features, w_i is a weight for feature *i*, and δ estimates the difference between the two instances' values at the *i*th feature. We employed the simple Overlap function that sets $\delta(x_i, y_i) = 0$ if $x_i = y_i$, and $\delta(x_i, y_i) = 1$ if $x_i \neq y_i$. The weight (importance) of a feature *i*, w_i, is estimated in our experiments by computing its gain ratio GR_i (Quinlan, 1993). These settings (k = 1, Overlap function, gain ratio feature weighting) are the default settings of the TiMBL software[1] (Daelemans et al., 2004) that we used to emulate memory-based learning, which we henceforth refer to as MBL.

2.3. Memory-based letter–phoneme conversion

MBL has been applied to word phonemization successfully in the past (Stanfill and Waltz, 1986; Van den Bosch and Daelemans, 1993; Van den Bosch et al., 1996; Daelemans and Van den Bosch, 2001), also to Dutch (Busser, 1998), and is known to perform at state-of-the-art performance levels as needed in speech synthesis technology. It has also been shown to outperform artificial neural networks (Stanfill and Waltz, 1986) and decision-tree methods (Daelemans et al., 1999). Many published machine-learning methods, including the one adopted here, make use of a deconstruction of the full word phonemization task into letter classification tasks: for each letter in a wordform, given a fixed "window" of wordform context, the classifier's task is to determine the phoneme label that this letter maps to. Table 1 displays example instances and their phoneme label classifications

[1] We used TiMBL version 5.1 for our experiments.

generated on the basis of the sample word <boeking> `booking'. Windows are generated spanning five left and right neighbor letters (or space characters when the window extends beyond the wordform). For example, the first instance in Table 1, "_ _ _ _ _ b o e k i n", maps to class label [b].

The spelling and the phonemic transcription of a word often differ in length, such as in the example <boeking> - [bukɪŋ]. Our windowed example approach demands, however, that the two representations be of equal length, so that each individual grapheme (one or more letters that are jointly mapped to one phoneme) can be mapped to a single phonemic symbol: <b> maps to [b], <oe> maps to [u], <k> maps to [k], <i> maps to [ɪ], and <ng> maps to [ŋ]. Our algorithm solves the letter–grapheme alignment problem fully automatically by aligning letters to phonemes; it does so by adding null phonemes in such a way that letters or strings of letters are consistently associated with the same phonemic symbols (e.g. <boeking> - [b̤-uk̤ɪ-ŋ], where the hyphen depicts a phonemic null). The actual choice which letter of a digraph becomes aligned to the null phoneme is a random but consistent choice made by the expectation–maximization (EM) algorithm (Dempster et al., 1977), which is employed to create an optimized letter-phoneme probability matrix automatically (for details, cf. Daelemans & Van den Bosch, 2001). Using the same mechanism, EM is used to insert graphemic nulls in those (relatively less frequent) cases in which the written form has fewer letters than the phonemic transcription has phonemes; e.g., <taxi> – [tɑksi]. Since graphemic nulls do not exist in spelling, they are implicitly encoded by mapping letters such as the <x> in <taxi> to diphones such as [ks].

Encoding the 293,825 unique phonemizations in windowed examples, a large database is produced containing 3,181,345 examples, each mapping to a phonemic label, or class. 205 unique classes occur, due to the existence of the double phonemes introduced with the EM-based alignment of words to phonemic transcriptions; for example, combinations of glides and vowels such as [ja] (in the pronunciation of <piano>) or [wo] (in the pronunciation of <duo>). The three most frequent classes are the phonemic null (13%), the [ə] (11%), and the [ʀ] (8%).

Example number	Left context	Focus letter	Right context	Class

1	_	_	_	_	_	b	o	e	k	i	n	[b]
2	_	_	_	_	b	o	e	k	i	n	g	-
3	_	_	_	b	o	e	k	i	n	g	_	[u]
4	_	_	b	o	e	k	i	n	g	_	_	[k]
5	_	b	o	e	k	i	n	g	_	_	_	[ɪ]
6	b	o	e	k	i	n	g	_	_	_	_	-
7	o	e	k	i	n	g	_	_	_	_	_	[ŋ]

Table 1. Examples generated for the word phonemization task, from the word-phonemization pair <boeking> - [bukɪŋ], aligned as [b̤-ukɪ̤-ŋ].

2.4. Memory-based morphological analysis

Memory-based learning has been proposed as a "single-level" approach to morphological analysis (Van den Bosch et al., 1996; Van den Bosch & Daelemans, 1999) (but see Clark, 2002 for an argument that the memory-based approach still involves non-trivial post-processing). To generate examples for the memory-based learner, each wordform and its associated analysis according to CELEX is converted into task instances using the windowing method exemplified in the previous subsection on word phonemization. Windowing transforms each wordform into as many instances as it has letters.

Example number	Left context					Focus letter	Right context					Class
1	_	_	_	_	_	a	b	n	o	r	m	A
2	_	_	_	_	a	b	n	o	r	m	a	0
3	_	_	_	a	b	n	o	r	m	a	l	0
4	_	_	a	b	n	o	r	m	a	l	i	0
5	_	a	b	n	o	r	m	a	l	i	t	0
6	a	b	n	o	r	m	a	l	i	t	e	0
7	b	n	o	r	m	a	l	i	t	e	i	0
8	n	o	r	m	a	l	i	t	e	i	t	0+Da
9	o	r	m	a	l	i	t	e	i	t	e	A_→N
10	r	m	a	l	i	t	e	i	t	e	n	0
11	m	a	l	i	t	e	i	t	e	n	_	0
12	a	l	i	t	e	i	t	e	n	_	_	0
13	l	i	t	e	i	t	e	n	_	_	_	0
14	i	t	e	i	t	e	n	_	_	_	_	plural
15	t	e	i	t	e	n	_	_	_	_	_	0

Table 2. Examples with morphological analysis classifications derived from the example word <abnormaliteiten> `abnormalities'. Each example focuses on one letter, and again includes a fixed number of five left and right neighbor letters.

To illustrate the construction of instances, Table 2 displays the 15 examples derived from the word <abnormaliteiten> `abnormalities' and their associated classes. The class of the first instance is A, which signifies that the morpheme starting in <a> is an adjective (A). The class of the eighth instance, 0+Da, indicates that at that position no segment starts (0), but that an <a> was deleted at that position (+Da, "delete a" here); this is due to the fact that the <l>, originally in word-final position, now is the onset of the following syllable, leaving the syllable <maa> open; in an open syllable, the convention in Dutch spelling is to represent long vowels with a single vowel letter. Next to deletions, insertions (+I) and replacements (+R, with a deletion and an insertion argument) can also occur. Together these two classification labels code that the first morpheme is the adjective <abnormaal> `abnormal'. The second morpheme, the suffix <iteit>, has class A_→N. This complex tag, which is in fact a rewrite rule, indicates that when <iteit> attaches right to an adjective (encoded by A_), the new combination becomes a noun (→N). Rewrite rule class labels occur exclusively with suffixes that do not have a part-of-speech tag of their own, but rather seek an attachment to form a complex morpheme with the part-of-speech tag. Finally, the third morpheme is <en>, which is a plural inflection that by definition attaches to a noun.

When a wordform is listed in CELEX as having more than one possible morphological labeling (e.g., a morpheme may be N or V, the inflection <-en> may be plural for nouns or infinitive for verbs), these labels are joined into ambiguous classes (N/V) and the first generated example is labeled with this ambiguous class. Ambiguity in syntactic and inflectional tags occurs in 3.6% of all morphemes in our CELEX data.

Encoding the data this way, a sizable data set of 3,179,383 windowed examples is generated; 2,738 different class labels occur. The most frequently occurring class label is 0, occurring in 68.8% of all instances. The three most frequent non-null labels are N (start of noun stem, 6.9%), V (start of verb stem, 3.6%), and plural (start of plural inflection, 1.6%). Many class labels combine a syntactic or inflectional tag with a spelling change, and generally have a low frequency.

2.5. Experimental baseline results for current Dutch

The purpose of computing a baseline score of the complexity of Dutch word phonemization and morphological analysis is to enable a clear comparison with the three alternative spellings for Dutch explored in the subsequent sections to the current state of Dutch. Based on the hypothesis that the success attained by an artificial learner in learning both of these tasks reflects the complexity of each task, the success of learning can be quantified by subjecting the learner to carefully designed training and testing experiments. It is important that the testing material does not overlap with the training set, as the MBL learner used in these experiments would

recognize any test word it had been trained to phonemize or morphologically analyze, and simply reproduce the correct pronunciation or analysis it has rote-learned. Consequently, a sensible test is composed of words the MBL learner has not been confronted with in training, and which forces it to generalize on the basis of its memory of example pronunciations or morphological analyses.

Another relevant choice in the composition of training and test data is to use word types or tokens. A token-based training and test set would optimally be drawn from a very large corpus of text, or would be sampled from a word type list where the draw is weighted by the relative frequencies of all types. Yet, this violates the formerly stated goal of focusing the performance on unseen words; any two samples of word tokens will contain a large overlap in words, and would lead to perfect phonemizations and morphological analyses of all words occurring in both sets, rote-learned by the memory-based learner. Since perfect reproduction does not reflect any interesting generalization performance, any token-based experiment will only poorly stress-test the generalization performance of the learner. The logical consequence is to draw at least the test data from a word type list that does not overlap with the training material.

To approximate this desired situation, the CELEX type list is split systematically using ten-fold cross-validation (Weiss & Kulikowski, 1991), a standard machine-learning methodology in which one dataset of examples is split in ten 10% subsets. Subsequently, ten experiments are performed in each of which nine 10% subsets are concatenated to form a 90% training set, on which the learner is trained, and in which the one held-out 10% test set is used for testing. This means that all word types in CELEX occur as a test word once, and are used nine times as one of the training words in the other nine experiments. Note that most words in CELEX are low-frequent words, which is typical for most unseen words when a list as CELEX is considered as a representative "known" words list. Also as any typical low-frequency word in any text, most words in the lexical database are morphologically complex. The average word length of CELEX word types is about eleven characters.

Evaluation of the two tasks was performed using the evaluation argued to be the best for each of the two tasks in the literature:

- Word phonemization is evaluated on the percentage of correctly phonemized test words. A word is phonemized correctly if all of its phonemes have been correctly generated by the classifier. Damper argues strongly for this word-level score as being the most relevant, in contrast to letter or phoneme-level scores (e.g., Damper and Eastmond, 1997).

- Morphological analysis is evaluated on the harmonic mean of the precision and recall of identifying morphemes (Van den Bosch and

Daelemans, 1999). An identification is counted as correct if the morpheme's boundaries are correctly identified and it is tagged correctly as an inflectional morpheme or a non-inflectional morpheme (i.e. a stem or an affix). Precision is the percentage of morphemes identified by the analyzer that are indeed morphemes in the target analysis; recall is the percentage of morphemes in the target analysis that are also predicted by the analyzer. Precision and recall can be merged in a single F-score, which is their harmonic mean: $F = \frac{2 \cdot precision \cdot recall}{precision + recall}$ (Van Rijsbergen, 1979).

The average percentage of correctly phonemized words over the ten 10-fold cross-validation experiments is 95.53%, with a standard deviation of 0.12. The average F-score on correctly identified morphemes over the 10 folds is 90.01%, with a standard deviation of 0.10. The low standard deviations indicate quite stable performance on different samples of unseen words.

Considering this, these two baseline scores represent the quite reasonable state of "current" Dutch with respect to the artificial learner's success in learning to generate phonological and morphological mappings based on spelling. Phonemization errors involve confusion with the letter <e>, often correlated with morphological ambiguities. For example, <pipetteren>, which ends in [-ɛrə],is rather similar to <etteren> which ends in [-ərə]; if <pipetteren> is in the test set, it incorrectly receives <etteren>'s pronunciation, and vice versa. A second major group of errors is due to ambiguities with loan word spellings. For example, <chansonnier> ought to have the French pronunciation, but there are many neighbor words ending in <-ier> with an [-ir] pronunciation. A third major group consists of errors with glides, partly induced by artefacts of the automatic EM-based alignment; for example, <vormvariant> `form variant' should be pronounced as [vɔʀmvaʀijɑnt], but is realized as [vɔʀmvaʀjɑnt].

Errors made by the learner in the morphological analysis task vary widely, but again three major classes become apparent when inspecting the predictions. First, certain long-distance dependencies are missed with long past participles; the prefix inflection is recognized, but the suffix inflection is mistaken for another verb inflection. Second, several segmentations between stems in compounds are missed. Third, analogous to the third group of word phonemization errors, artefact inconsistencies in the analyses of long compounds in CELEX crop up as mutually incorrect nearest neighbors if the inconsistent neighbors are divided over training and test set.

Given our argumentation that the Dutch spelling system represents a certain balance between adhering to phonological and morphological principles, it may be helpful to view the current and possible alternative states of Dutch spelling as points in a two-dimensional space, of which the axes represent the phonological and morphological learnability results. The ideal spelling lies

at the point in the upper right corner of this space, with 100% correct phonemic mapping, and an F-score of 100% on correctly identified morphemes. Figure 2 visualizes the upper right corner of the space starting from 80% in both evaluation metrics. The position of "current" Dutch is marked with the label "current" (the labels "morphemic" and "schwa" are explained and discussed in subsequent sections). The curved line is an F-isoline; when the harmonic mean (i.e., F-score) of two evaluation metrics would be taken, this isoline represents all points for which the harmonic mean is 90%. The isoline can be seen as a height isoline in a topographic map, where any improvement is one that goes uphill in the direction of the upper right corner.

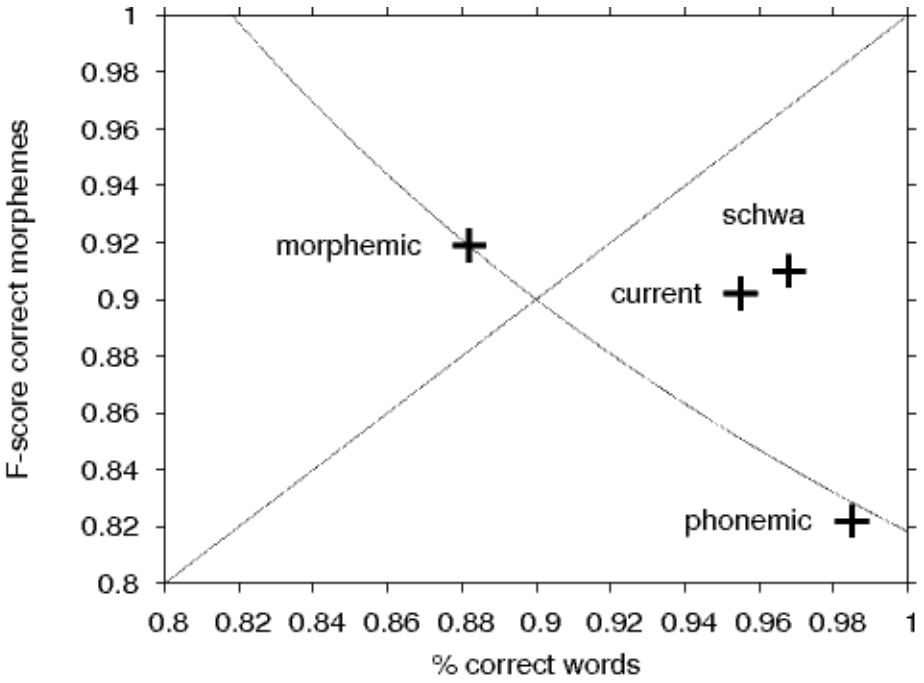

Figure 2. Upper right corner of the two-dimensional spelling space, in which the outcomes of all four experiments of this article are plotted. The curved lines are the F-isolines for F = 0.9 (upper) and F=0.8 (lower).

3. Artificial regularization of letter–phoneme correspondences

Our first experimental manipulation is to make the spelling of Dutch more phonological (or rather, more phonemic) which would likely make the word phonemization task very easy to learn. The letter–phoneme mappings of all words are regularized simply by spelling out their SAMPA transcriptions while using the letters of the normal alphabet, according to these simple rules:

1. Spell out the schwa as an <e>.
2. Spell out all short vowels using single vowel letters.

3. Spell out all long vowels using geminated vowel letters and the other usual two-letter vowel letter conventions of Dutch (e.g. <ie> for [i]).
4. Wipe out the historical ambiguity between <ei> and <ij>, and write only <ij>; analogously, write only <au> instead of <au> or <ou>, and only <s> instead of word-final <sch>.
5. Spell out glides.

For example, the word <piano>, pronounced [pijano], is spelled <piejaanoo> in the phonemic spelling. The word <aanlokkelijk> `attractive', pronounced [anlɔkələk], is re-spelled <aanlokelek>. Note that in the latter example the first <k> is not geminated, as in the original, which would normally signal that the <o> is short; but in this spelling, the single <o> already signals that. Also note that spelling the schwa with <e> actually leaves some ambiguity, since an <e> can also be pronounced as [ɛ] in this phonemic spelling.

After rewriting all words, EM is used to re-align spelling to phonology; although more phonemic nulls need to be inserted (words tend to become longer on average due to the double vowels), this alignment converges to a very high joint probability. While rewriting each word's spelling, its corresponding morphological structure is adapted along the changes, so that when a word's stem becomes one letter longer (such as in <baazen> `bosses', the phonemic spelling variant of <bazen>), the segmentation information is shifted one position as well.

Again two ten-fold cross-validation experiments were performed, yielding the following results. The word phonemization task is indeed learned at a higher level of performance; on unseen words, an average correctness level of 98.52 (standard deviation 0.07) correctly phonemized words is observed. Indeed, the few word phonemization errors made involve errors on words with the letter <e>. On the other hand, the F-score on correctly identified morphemes drops to 83.53 (standard deviation 0.17). In the spelling space visualized in Figure 2, this point is marked "phonemic". This point is far less harmonic than the "current" state of Dutch, and more distant from the perfect upper right corner. The overall harmony of this spelling has clearly deteriorated from the perspective of learnability of the two dimensions.

The increased error in morphological analysis occurs largely because the classifier refrains from predicting segmentations in many instances. Many wordforms become more ambiguous in the phonemized spelling. For example, the word <soliditeit> `solidity' changes to <sooliedietijt>. The suffix <-iteit> (-ity) now becomes confused with the stem <tijd> (time, as if in `solidy time'); the confused morphological analyzer decides not to split before <itijt> nor before <tijt>.

4. Artificial regularization of morphological structure

The second manipulation of Dutch spelling aims to preserve the spelling of all morphemes in the overall surface spelling of words: in all words a morpheme occurs in, it should be spelled the same. All inflections and derivational suffixes are spelled in their full form, and stems take the uninflected form, and in case of nouns and verbs, the singular uninflected form; a choice also made in the morphological analyses of CELEX. To generate the data for the experiment, a morphological spelling of each word is generated by simply concatenating the stems and inflectional or derivational morphemes listed in the morphological analysis provided by CELEX. For example, the word <huizen> `houses' is spelled as <huisen>, a concatenation of the stem <huis> and <en>. The word <bruggen> `bridges' is re-spelled as <brugen>, without the usual gemination marking that <u> is a short vowel. The example word from Table 2, <abnormaliteiten>, segmented according to CELEX as [ab][normaal][iteit][en], is spelled as <abnormaaliteiten>, differing in the double <aa> versus the single <a>.

As expected, this manipulation renders the morphological analysis task more learnable. In a ten-fold cross-validation experiment an F-score of 91.94 (standard deviation 0.05) is attained, which is 1.8% better than the 90.01% F-score of "current" Dutch. On the other hand, the learnability of word phonemization drops to an average score of 88.20% correctly phonemized words (standard deviation 0.16). In Figure 2 the position of this morphemic spelling is marked with "morphemic". Again, this point in spelling space is more distant to the upper right corner than the "current" state. The artificial regularization of morphological structure leads to an overall deterioration of the learnability of word phonemization, much like the artificial regularization of the relation between letters and phonemes lead to an overall decrease in the capability of the artificial learner to generate morphological analyses.

The deterioration of errors in word phonemization can be attributed largely to the masking of information needed for phonemic mappings due to the strict retention of the spelling of stems. The word <flatteuze> `flattering' is re-spelled as <flatteuse>, restoring the <s> of the uninflected form <flatteus>, causing an incorrect pronunciation with an [s]. A similar case is <verstijving> `stiffening', re-spelled as <verstijfing>, of which the <f> is pronounced as [f].

5. Optimizing spelling space: The schwa as a letter

In Section 3 it was already mentioned that the schwa ([ə]) is involved in the only ambiguity left in the phonemic spelling variant. In Dutch, the schwa is often spelled as an <e>, but this letter can also map to [e] and [ɛ]. It is quite essential to note that the pronunciation of the <e> is for a major part

related to morphological structure. In most of the frequently occurring inflectional morphemes with an <e>, such as <-e>, <-en>, <-end>, <-ende>, <be->, <ge->, the <e> is pronounced as a schwa. This fully disambiguates the pronunciation of, for example, <bekeren> `to convert'; a correct morphological analysis would lead to the determination that <be-> and <-en> are pronounced with schwas, while the stem part <ker> (from <keer> `turn') is pronounced with an [e], resulting in [bəkerə].

The "current" spelling of Dutch was adapted by replacing all vowels with the keyboard character <@> (the at sign) when mapping to a schwa in the aligned phonemic representation. Subsequently two ten-fold cross-validation experiments were performed on the two learning tasks, resulting in a score of 96.82% correctly phonemized words (standard deviation 0.13), 1.3% better than "current" Dutch. On morphological analysis an F-score of 90.89% (standard deviation 0.05) was obtained, 0.9% better than "current" Dutch. This means the "schwa" manipulation, also plotted in Figure 2, represents a genuine improvement of the spelling, incurring improvements in both dimensions.

6. Discussion

Through learnability experiments with an artificial learner, evidence is gathered for a balance in the way phonology and morphology determine the current spelling of Dutch words. It is a well-known fact, not exclusively for Dutch but rather for all European alphabetic writing systems, that the phonological and morphological principles are in competition with each other. Sometimes the phonological principle deletes vowels, geminates consonants or changes <v> into <f> or <z> into <s> in final devoicing. At other occasions, the morphological principle introduces phonetically odd geminations and holds on to letters that are not pronounced as would be expected in a regular letter-phoneme mapping. Still, they do so quite in balance. From these experiments it can be concluded that the resulting system, current Dutch spelling, is quite balanced between adhering to the phonological and morphological principles – in terms of learnability of phonological and morphological mappings. This is a non-trivial observation – certainly it is not a case for making Dutch more phonologically regular. As showed in the experiments with the phonologically regularized variant of Dutch spelling, word phonemization can indeed be reduced to a trivial task, but the point to make here is that at the same time morphological analysis becomes a much harder task. Analogously, making the spelling adhere fully to the morphological principle makes the word phonemization task considerably harder.

By pursuing the reasoning that the schwa on the one hand causes confusion with phonemic mappings of the letter <e>, while on the other hand it is the common pronunciation of <e>s in many inflectional morphemes and affixes,

the assumption was made that it might be fruitful to use the schwa as a letter (the <@> keyboard character was used) in positions in which some vowel, usually the <e>, is actually pronounced as a schwa. This variant caused minor but interesting improvements in both dimensions. Set aside the question whether the schwa would ever be accepted as a letter, it shows that the balance in "current" Dutch is not the best balance that Dutch could theoretically achieve.

This contribution has presented three measurements in a space in which many more points are possible; many more variants of Dutch spelling could be devised that would try to balance some aspect of morphological regularity with phonemization regularity better than the two extremes chosen to be emulated here. The extremes were based on simplistic reasonings that deliberately ignored any issue that would counter the damage incurred to the other dimension. Aside from the "schwa" experiment, many more subtle and linguistically-motivated experiments could be performed. For instance, a more careful "morphological" variant of Dutch could retain the gemination of consonants also in the singular forms, as in German orthography (e.g., to respell the singular form of <brug> as <brugg>, due to the plural <bruggen>), which keep the pronunciation of the vowel in the stem regularly predictable.

The presented system in fact offers a straightforward computational tool to explore such variations and subtle optimizations. I refrain, however, from claiming that the method presented here is a full-blown solution to spelling reform. There are many cognitive, perceptive, and emotional aspects involved in spelling that are obviously not measured in these experiments. Rather, the proposed method of using an artificial learner and measuring its performance in a two-dimensional space, in which changes can be measured according to their harmony and progress towards the perfect spelling, forms a useful testbed for spelling reformers to assist them in decision making, and to signal unforeseen consequences of certain simplifications in one dimension to other dimensions.

Using an artificial learner abstracts from, and is by no means meant to be a realistic model of human learning. Yet, there is an implied relation between the learnability and complexity measurements taken in these experiments, and the ease with which human learners, be it first-language-learning children or adult second-language learners, learn to read, pronounce, and write Dutch words. This line of reasoning finds its roots in psycholinguistic modeling work taking the perspective of the reader, the speaker, and the writer, both at the phonological level (Glushko, 1979; Plaut et al., 1996) and the morphological level (Plaut and Gonnerman, 2000), and work focusing on letter-to-sound processing (Geudens and Sandra, 1999; Diependaele, Sandra, and Grainger, 2005). The connectionist models used in some of the modeling studies cited here are largely equivalent to the memory-based models used in the present study. However, MBL models do not directly predict reaction times, word image effects, or any longer or broader memory

and learning effects seen in children or second-language learners. The current experiments also do not reflect the amount and type of material a child becomes gradually acquainted with over time. It would be an interesting point for further research to account for these aspects and connect with psycholinguistic work on learning a spelling system (e.g. Gillis & Sandra, 2000; Sandra, 2003), as well as with work focusing on the effect of actual spelling changes, such as performed by Schreuder et al. (1998) and Neijt et al. (2004), in the wake of the 1995 Dutch spelling reform.

A second, related point for further research is to distinguish between effects of spelling change on reading versus writing. For a reader who is pronouncing text, the difference between <ei> and <ij> is trivial; both are pronounced the same way in all contexts. For the writer, however, the choice between the two is hard, since it is lexically determined; simplifying the choice by always choosing one spelling (as done in the phonological spelling experiment here) would arguably alleviate the writer's task, but would not change the ease of the pronunciation task. In general, spelling changes should improve both tasks, and it would be desirable to decompose the lump sum measurement taken in this study, into separate measurements of the different degrees in which spelling changes affect reading or writing. One option is to investigate the complexity and learnability of phoneme-to-grapheme conversion (Decadt et al., 2002).

A third issue for further research is the expansion of the space to include the letter-grapheme complexity dimension. As argued by Van den Bosch et al. (1994), this dimension is not necessarily correlated with the complexity of grapheme–phoneme conversion. For instance, the French spelling system has rather ambiguous vowel letter combinations such as <au> and <eau> that, once graphemically delineated in a word, have a very regular phonemization. This problem was hidden in the current study by pre-aligning our phonemization data using EM, but it might be relevant to measure letter–grapheme mapping in isolation. Another candidate dimension is word stress, which is not symmetrically correlated to word phonemization, and has interesting relations with morphological structure in terms of learnability (Busser, 1998).

Acknowledgements

This research was funded by NWO, The Netherlands Organization for Scientific Research. The author wishes to thank Walter Daelemans, Martin Neef, and two anonymous reviewers for their valuable suggestions.

References

Aha, D. W., Kibler, D., & Albert, M. (1991). Instance-based learning algorithms. Machine Learning, 6, 37–66.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). The CELEX lexical data base on CD-ROM. Philadelphia, PA: Linguistic Data Consortium.

Busser, G. J. (1998). TreeTalk-D: a machine learning approach to Dutch word pronunciation. Proceedings of the Text, Speech, and Dialogue Conference (pp. 3–8).

Clark, A. (2002). Memory-based learning of morphology with stochastic transducers. Proceedings of the 40th Meeting of the Association for Computational Linguistics (pp. 513–520). New Brunswick, NJ: ACL.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. Institute of Electrical and Electronics Engineers Transactions on Information Theory, 13, 21–27.

Daelemans, W., & Van den Bosch, A. (2001). Treetalk: Memory-based word phonemisation. In R. I. Damper (Ed.), Data-driven techniques in speech synthesis, 149–172. Dordrecht: Kluwer Academic Publishers.

Daelemans, W., Van den Bosch, A., & Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. Machine Learning, Special issue on Natural Language Learning, 34, 11–41.

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (2004). TiMBL: Tilburg memory based learner, version 5.1, reference guide (Technical Report ILK 04-02). ILK Research Group, Tilburg University.

Daelemans, W., and Van den Bosch, A. (2005). Memory-based language processing. Cambridge, UK: Cambridge University Press.

Damper, R. I., & Eastmond, J. F. G. (1997). Pronunciation by analogy: impact of implementational choices on performance. Language and Speech, 40, 1–23.

Decadt, B., Duchateau, J., Daelemans, W., and Wambacq, P. (2002). Memory-based phoneme-to-grapheme conversion. In: M. Theune, A. Nijholt and H. Hondrop (Eds), Proceedings of the Twelfth Meeting of Computational Linguistics in the Netherlands (CLIN 2001). Amsterdam: Rodopi, pp. 47-61.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 39, 1–38.

Diependaele, K., Sandra, D., and Grainger, J. (2005). Masked cross-modal morphological priming: Unravelling morpho-orthographic and morpho-semantic influences in early word recognition.
Language and Cognitive Processes, 20(1-2), 75-114.

Geudens, A., & Sandra, D. (1999). Onsets and rimes in a phonologically transparent orthography: Differences between good and poor beginning readers of Dutch. Brain and Language, 68, 284-290.

Gillis, S., & Sandra, D. (2000). The influence of spelling rules on phonology. In M. Beers (Ed.), From sound to sentence: Studies on first language acquisition, 43–55.

Groningen: Centre for Language and Cognition.

Glushko, R. J. (1979). The organisation and activation of orthographic knowledge in reading aloud. Journal of Experimental Psychology: Human Perception and Performance, 5, 647–691.

Neijt, A., Schreuder, R., & Baayen, R. H. (2004). Seven years later: The effect of spelling on interpretation. In L. Cornips and J. Doetjes (Eds.), Linguistics in the Netherlands 2004, 134–145. Amsterdam: Benjamins.

Plaut, D., McClelland, J., Seidenberg, M., and Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review, 103, 56-115.

Plaut, D. C. and Gonnerman, L. M. (2000) Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? Language and Cognitive Processes, 15, 445-485.

Quinlan, J. (1993). c4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.

Raible, W. (1991). Zur Entwicklung von Alphabetschrift-Systemen. No. 1 in Abhandlungen der Heidelberger Akademie der Wissenschaften, Philosophisch-historische Klasse. Heidelberg: Winter.

Sandra, D. (2003). Spelling errors with a view on the mental lexicon: frequency and proximity effects in misspelling homophonous regular verb forms in Dutch and French. In R. H. Baayen (Ed.), Trends in linguistics: morphological structure in language processing, 485–514. Den Haag: De Gruyter.

Schreuder, R., Neijt, A., Van derWeide, F., & Baayen, R. H. (1998). Regular plurals in Dutch compounds: linking graphemes or morphemes? Language and Cognitive Processes, 13, 551–573.

Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. Communications of the ACM, 29, 1213–1228.

Van den Bosch, A., Content, A., Daelemans, W., & De Gelder, B. (1994). Measuring the complexity of writing systems. Journal of Quantitative Linguistics, 1:3, 178-188.

Van den Bosch, A., & Daelemans, W. (1993). Data-oriented methods for grapheme-to-phoneme conversion. Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (pp. 45-53).

Van den Bosch, A., & Daelemans, W. (1999). Memory-based morphological analysis. Proceedings of the 37th Annual Meeting of the ACL (pp. 285–292). San Francisco, CA: Morgan Kaufmann.

Van den Bosch, A., Daelemans, W., & Weijters, A. (1996). Morphological analysis as classification: an inductive-learning approach. Proceedings of the Second International Conference on New Methods in Natural Language Processing, NeMLaP-2, Ankara, Turkey (pp. 79–89).

Van Rijsbergen, C. (1979). Information retrieval. London: Buttersworth.

Weiss, S., & Kulikowski, C. (1991). Computer systems that learn. San Mateo, CA: Morgan Kaufmann.