# $S$-restricted monotone alignments

**Steffen Eger**
Faculty of Economics
Goethe Universität Frankfurt
Grüneburg-Platz 1, 60323 Frankfurt am Main, Germany
steffen.eger@yahoo.com

## Abstract

We present an alignment algorithm for monotone many-to-many alignments, which are relevant e.g. in the field of grapheme-to-phoneme conversion (G2P). Moreover, we specify the size of the search space for monotone many-to-many alignments in G2P, which indicates that exhaustive enumeration is generally possible, so that some limitations of our approach can easily be overcome. Finally, we present a decoding scheme, within the monotone many-to-many alignment paradigm, that relates the decoding problem to restricted integer compositions and that is, putatively, superior to alternatives suggested in the literature.

## 1 Introduction

Grapheme-to-phoneme conversion (G2P) is the problem of transducing, or converting, a grapheme, or letter, string **x** over an alphabet $\Sigma_x$ into a phoneme string **y** over an alphabet $\Sigma_y$. An important first step thereby is finding *alignments* between grapheme and phoneme strings in training data. The classical alignment paradigm has presupposed alignments that were

(i) *one-to-one* or *one-to-zero*; i.e. one grapheme character is mapped to at most one phoneme character; this assumption has probably been a relic of both the traditional assumptions in machine translation (Brown et al. 1990) and in biological sequence alignment (Needleman and Wunsch, 1970). In the field of G2P such alignment models are sometimes also called $\epsilon$-scattering models (Black et al., 1998).

(ii) *monotone*, that is, the order between characters in grapheme and phoneme strings is preserved.

It is clear that, despite its benefits, the classical alignment paradigm has a couple of limitations; in particular, it may be unable to explain certain grapheme-phoneme sequence pairs, a.o. those where the length of the phoneme string is greater than the length of the grapheme string such as in

exact igzækt

where **x** has length 5 and **y** has length 6. In the same context, even if an input pair can be explained, the one-to-one or one-to-zero assumption may lead to alignments that, linguistically, seem nonsensical, such as

| p | h | o | e | n | i | x |
|---|---|---|---|---|---|---|
| f | – | iː | n | i | k | s |

where the reader may verify that, no matter where the $\epsilon$ is inserted, some associations will always appear unmotivated. Moreover, monotonicity appears in some cases violated as well, such as in the following,

centre sentər

where it seems, linguistically, that the letter character *r* corresponds to phonemic r and graphemic word final *e* corresponds to ə.

Fortunately, better alignment models have been suggested to overcome these problems. For example, Jiampojamarn et al. (2007) and Jiampojamarn and Kondrak (2010) suggest 'many-to-many' alignment models that address issue (i) above. Similar ideas were already present in (Baldwin and Tanaka, 2000), (Galescu and Allen, 2001) and (Taylor, 2005). Bisani and Ney (2008) likewise propose many-to-many alignment models; more precisely, their idea is to *segment* grapheme-phoneme pairs into non-overlapping parts ('co-segmentation'), calling each segment a *graphone*, as in the following example, consisting of five graphones,

| ph | oe | n | i | x |
|---|---|---|---|---|
| f | iː | n | i | ks |

The purpose of the present paper is to introduce a very simple, flexible and general monotone many-to-many alignment algorithm (in Section 3) that competes with the approach suggested in Jiampojamarn et al. (2007). Thereby, our algorithm is an intuitive and straightforward generalization of the classical Needleman-Wunsch algorithm for (biological or linguistic) sequence alignment. Moreover, we explore simple and valuable extensions of the presented framework, likewise in Section 3, which may be useful e.g. to detect latent classes in alignments, similar to what has been done in e.g. Dreyer et al. (2008). We also mention limitations of our procedure, in Section 4, and discuss the naive brute-force approach, exhaustive enumeration, as an alternative; furthermore, by specifying the search space for monotone many-to-many alignments, we indicate that exhaustive enumeration appears generally a feasible option in G2P and related fields. Then, a second contribution of this work is to suggest an alternative decoding procedure when transducing strings $\mathbf{x}$ into strings $\mathbf{y}$, within the monotone many-to-many alignment paradigm (in Section 6.2). We thereby relate the decoding problem to restricted integer compositions, a field in mathematical combinatorics that has received increased attention in the last few years (cf. (Heubach and Mansour, 2004; Malandro, 2012)). Finally, we demonstrate the superiority of our approach by applying it to several data sets in Section 7.

It must be mentioned, generally, that we take G2P only as an (important) sample application of monotone many-to-many alignments, but that they clearly apply to other fields of natural language processing as well, such as transliteration, morphology/lemmatization, etc. and we thus also incorporate experiments on morphology data. Moreover, as indicated, we do not question the premise of monotonicity in the current work, but take it as a crucial assumption of our approach, leading to efficient algorithms. Still, 'local non-monotonicities' as exemplified above can certainly be adequately addressed within our framework, as should become clear from our illustrations below (e.g. with higher-order 'steps').

## 2 $S$-restricted paths and alignments

Consider the two-dimensional lattice $\mathbb{Z}^2$. In $\mathbb{Z}^2$, we call an ordered list of pairs $(\alpha_0, \beta_0) = (0,0), \ldots, (\alpha_k, \beta_k) = (m,n)$ a *path* from $(0,0)$ to $(m,n)$, and we call $(a_i, b_i) := (\alpha_i, \beta_i) - (\alpha_{i-1}, \beta_{i-1})$, $i = 1, \ldots, k$, *steps*. Moreover, we call a path $\lambda$ in the lattice $\mathbb{Z}^2$ from $(0,0)$ to $(m,n)$ *monotone* if all steps $(a,b)$ are non-negative, i.e. $a \geq 0$, $b \geq 0$, and we call the monotone path $\lambda$ $S$-restricted for a subset $S$ of $\mathbb{N}^2$ if all steps lie within $S$, i.e. $(a,b) \in S$.

Note that $S$-restricted monotone paths define (restricted) co-segmentations, or (a special class of) monotone alignments, between strings $\mathbf{x}$ and $\mathbf{y}$. For example, the two paths in Figure 1 correspond to the two monotone alignments between $\mathbf{x}$ = phoenix and $\mathbf{y}$ = fiːniks illustrated above. Thus, we identify $S$-restricted monotone paths with $S$-restricted monotone alignments in the sequel.

Moreover, note that the set and number of $S$-restricted monotone paths allow simple recursions. To illustrate, the number $T_S(m,n)$ of $S$-restricted monotone paths from $(0,0)$ to $(m,n)$ satisifies

$$T_S(m,n) = \sum_{(a,b) \in S} T_S(m-a, n-b), \qquad (1)$$

with initial condition $T_S(0,0) = 1$ and $T_S(m,n) = 0$ if $m < 0$ or $n < 0$. As will be seen in the next section, under certain assumptions, *optimal* monotone alignments (or, equiva-

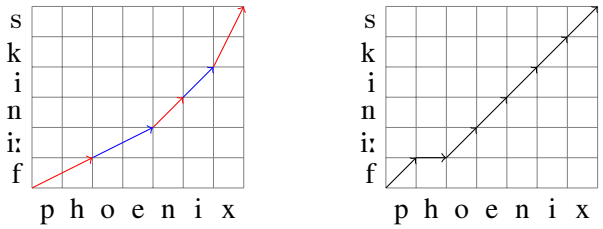


Figure 1: Monotone paths in two-dimensional lattices corresponding to the monotone alignments between $\mathbf{x}$ = phoenix and $\mathbf{y}$ = fiːniks given in Section 1. In the left lattice, we have arbitrarily (but suggestively) colored each step in either red or blue.

lently, paths) can be found via a very similar recursion.

## 3 Algorithm for $S$-restricted alignments

Let two strings $\mathbf{x} \in \Sigma_x^*$ and $\mathbf{y} \in \Sigma_y^*$ be given. Moreover, assume that a set $S$ of allowable steps is specified together with a real-valued similarity function $\mathsf{sim} : \Sigma_x^* \times \Sigma_y^* \to \mathbb{R}$ between characters of $\Sigma_x$ and $\Sigma_y$. Finally, assume that the *score* or value of an $S$-restricted monotone path $\lambda = (\alpha_0, \beta_0), \ldots, (\alpha_k, \beta_k)$ is defined additively linear in the similarity of the substrings of $\mathbf{x}$ and $\mathbf{y}$ corresponding to the steps $(a, b)$ taken, i.e.

$$\text{score}(\lambda) = \sum_{i=1}^{k} \mathsf{sim}(x_{\alpha_{i-1}+1}^{\alpha_i}, y_{\beta_{i-1}+1}^{\beta_i}), \qquad (2)$$

where by $x_{\alpha_{i-1}+1}^{\alpha_i}$ we denote the subsequence $x_{\alpha_{i-1}+1} \ldots x_{\alpha_i}$ of $\mathbf{x}$ and analogously for $\mathbf{y}$. Then it is not difficult to see that the problem of finding the path (alignment) with maximal score can be solved efficiently using a very similar (dynamic programming) recursion as in Eq. (1), which we outline in Algorithm 1. Moreover, this algorithm is obviously a straightforward generalization of the classical Needleman-Wunsch algorithm, which specifies $S$ as $\{(0,1),(1,0),(1,1)\}$.

Note, too, that in Algorithm 1 we include two additional quantities, not present in the original sequence alignment approach, namely, firstly, the 'quality' $\mathsf{q}$ of a step $(a, b)$, weighted by a factor $\gamma \in \mathbb{R}$. This quantity may be of practical importance in many situations. For example, if we specify $\mathsf{sim}$ as log-probability (see below), then Algorithm 1 has a 'built-in' tendency to substitute 'smaller', individually more likely steps $(a, b)$ by larger, less likely steps because in the latter case fewer negative numbers are added; if $\mathsf{sim}$ assigns strictly positive values, this relationship is reversed. We can counteract these biases by factoring in the *per se* quality of a given step. Also note that if $\mathsf{q}$ is added linearly, as we have specified, then the dynamic programming recursion is not violated.

**Algorithm 1** Gen. Needleman-Wunsch (GNW)

**procedure** GNW($x_1 \ldots x_m, y_1 \ldots y_n$; $S$, $\mathsf{sim}, \mathsf{q}, \mathsf{L}$)
  $M_{ij} \leftarrow 0$ for all $(i,j) \in \mathbb{Z}^2$ such that $i < 0$ or $j < 0$
  $M_{00} \leftarrow 1$
  **for** $i = 0 \ldots m$ **do**
    **for** $j = 0 \ldots n$ **do**
      **if** $(i,j) \neq (0,0)$ **then**
        $M_{ij} \leftarrow \max_{(a,b)^c \in S} \{M_{i-a,j-b} + \mathsf{sim}(x_{i-a+1}^i, y_{j-b+1}^j) + \gamma \mathsf{q}(a,b) + \chi \mathsf{L}\big((x_{i-a+1}^i, y_{j-b+1}^j), c\big)\}$
      **end if**
    **end for**
  **end for**
  **return** $M_{mn}$
**end procedure**

Secondly, we specify a function $\mathsf{L} : \big(\Sigma_x^* \times \Sigma_y^*\big) \times \mathsf{colors} \to \mathbb{R}$, where $\mathsf{colors}$ is a finite set of 'colors', that encodes the following idea. Assume that each step $(a, b) \in S$ appears in $C$, $C \in \mathbb{N}$, different 'colors', or states. Then, when taking step $(a, b)$ with color $c \in \mathsf{colors}$ (which we denote by the symbol $(a, b)^c$ in Algorithm 1), we assess the 'goodness' of this decision by the 'likelihood' $\mathsf{L}$ that the current subsequences of $\mathbf{x}$ and $\mathbf{y}$ selected by the step $(a, b)$ 'belong to'/'are of' color (or state) $c$. As will be seen below, this al-

lows to very conveniently identify (or postulate) 'latent classes' for character subsequences, while increasing the algorithm's running time only by a constant factor.

As to the similarity measure $\mathsf{sim}$ employed in Algorithm 1, a popular choice is to specify it as the (logarithm of the) joint probability of the pair $(u, v) \in \Sigma_x^* \times \Sigma_y^*$, but a multitude of alternatives is conceivable here such as the $\chi^2$ similarity, pointwise mutual information, etc. (see for instance the overview in Hoang et al. (2009)). Also note that if $\mathsf{sim}$ is e.g. defined as joint probability $\Pr(u, v)$ of the string pair $(u, v)$, then $\Pr(u, v)$ is usually initially unknown but can be iteratively estimated via application of Algorithm 1 and count estimates in an EM-like fashion (Dempster et al., 1977), see Algorithm 2. As concerns $\mathsf{q}$ and $\mathsf{L}$, we can likewise estimate them iteratively from data, specifying their abstract forms via any well-defined (goodness) measures. The associated coefficients $\gamma$ and $\chi$ can be optimized on a development set or set exogenously.

**Algorithm 2** (Hard) EM Training

**procedure** EM($\{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \ldots, N\}$; $S$, $T$, $\hat{\mathsf{sim}}_0, \hat{\mathsf{q}}_0, \hat{\mathsf{L}}_0$)
  $t \leftarrow 0$
  **while** $t < T$ **do**
    **for** $i = 1 \ldots N$ **do**
      $(\mathbf{x}_i^a, \mathbf{y}_i^a) \leftarrow \mathrm{GNW}(\mathbf{x}_i, \mathbf{y}_i; S, \hat{\mathsf{sim}}_t, \hat{\mathsf{q}}_t, \hat{\mathsf{L}}_t)$
    **end for**
    $\hat{\mathsf{sim}}_{t+1}, \hat{\mathsf{q}}_{t+1}, \hat{\mathsf{L}}_{t+1} \leftarrow f(\{\mathbf{x}_i^a, \mathbf{y}_i^a \mid i = 1, \ldots, N\})$
    $t \leftarrow t + 1$
  **end while**
**end procedure**

## 4 Exhaustive enumeration and alignments

In the last section, we have specified a polynomial time algorithm for solving the monotonic $S$-restricted string alignment problem, under the following restriction; namely, we defined the score of an alignment additively linear in the similarities of the involved subsequences. This, however, entails an independence assumption between successive aligned substrings that oftentimes does not seem justified in linguistic applications. If, on the contrary, we specified the score, $\mathrm{score}(\lambda)$, of an alignment $\lambda$ between strings $\mathbf{x}$ and $\mathbf{y}$ as e.g.

$$\sum_{i=1}^{k} \log \Pr\big((x_{\alpha_{i-1}+1}^{\alpha_i}, y_{\beta_{i-1}+1}^{\beta_i}) \mid (x_{\alpha_{i-2}+1}^{\alpha_{i-1}}, y_{\beta_{i-2}+1}^{\beta_{i-1}})\big)$$

(using joint probability as similarity measure) — this would correspond to a 'bigram scoring model' — then Algorithm 1 would not apply.

To address this issue, we suggest *exhaustive enumeration* as a possibly noteworthy alternative — enumerate *all* $S$-restricted monotone alignments between strings $\mathbf{x}$ and $\mathbf{y}$, score each of them individually, taking the one with maximal score. This brute-force approach is, despite its simplicity, the most general approach conceivable and works under all specifications of scoring functions. Its practical applicability relies on the sizes of the search spaces for $S$-restricted monotone alignments and on the lengths of the strings $\mathbf{x}$ and $\mathbf{y}$ involved.

We note the following here. By Eq. 1, for the choice $S = \{(1,1), (1,2), (1,3), (1,4), (2,1)\}$, a seemingly reasonable specification in the context of G2P (see below), the number $T_S(n, n)$ of $S$-restricted monotone alignments is given as (for explicit formulae, cf. (Eger, 2012))

$$1, 1, 3, 7, 16, 39, 95, 233, 572, 1406, 3479, 8647$$

for $n = 1, 2, \ldots, 12$ and e.g. $T_S(15, 15) = 134,913$. Moreover, for the distribution of letter string and phoneme string lengths we estimate Poisson distributions (Wimmer et al., 1994) with parameters $\mu \in \mathbb{R}$ as listed in Table 1 for the German Celex (Baayen et al., 1996), French Brulex (Content et al., 1990) and English Celex datasets, as used in Section 7. As the table and the above numbers show, there are on average only a few hundred or few thousand possible monotone many-to-many alignments between grapheme and phoneme string pairs, for which exhaustive enumeration appears, thus, quite feasible; moreover, given enough data, it usually does not harm much to exclude a few string pairs, for which alignment numbers are too large.

## 5 Choice of $S$

Choice of the set of steps $S$ is a question of *model selection*, cf. (Zucchini, 2000). Several ap-

| Dataset | $\mu_G$ | $\mu_P$ | $P_{[G>15]}$ | $P_{[P>15]}$ |
|---|---|---|---|---|
| German-Celex | 9.98 | 8.67 | 4.80% | 1.62% |
| French-Brulex | 8.49 | 6.71 | 1.36% | 0.15% |
| English-Celex | 8.21 | 7.39 | 1.03% | 0.40% |

Table 1: Avg. grapheme and phoneme string lengths in resp. data set, and probabilities that lengths exceed 15.

proaches are conceivable here. First, for a given domain of application one might specify a possibly 'large' set of steps $\Omega$ capturing a preferably comprehensive class of alignment phenomena in the domain. This may not be the best option because it may provide Algorithm 1 with too many 'degrees of freedom', allowing it to settle in unfavorable local optima. A better, but potentially very costly, alternative is to exhaustively enumerate all possible subsets $S$ of $\Omega$, apply Algorithm 1 and/or Algorithm 2, and evaluate the quality of the resulting alignments with any choice of suitable measures such as alignment entropy (Pervouchine et al., 2009), average log-likelihood, Akaike's information criterion (Akaike, 1974) or the like. Another possibility would be to use a comprehensive $\Omega$, but to penalize unlikely steps, which could be achieved by setting $\gamma$ in Algorithm 1 to a 'large' real number and then, in subsequent runs, employ the remaining steps $S \subseteq \Omega$; we outline this approach in Section 7.

Sometimes, specific knowledge about a particular domain of application may be helpful, too. For example, in the field of G2P, we would expect most associations in alignments to be of the type $M$-to-1, i.e. one or several graphemes encode a single phoneme. This is because it seems reasonable to assume that the number of phonetic units used in language communities typically exceeds the number of units in alphabetic writing systems — 26 in the case of the Latin alphabet — so that one or several letters must be employed to represent a single phoneme. There may be 1-to-$N$ or even $M$-to-$N$ relationships but we would consider these exceptions. In the current work, we choose $S = \{(1,1),(2,1),(3,1),(4,1),(1,2)\}$ for G2P data sets, and for the morphology data sets we either adopt from (Eger, 2012) or use a comprehensive $\Omega$ with 'largest' step $(2,2)$.

# 6 Decoding

## 6.1 Training a string transduction model

We first generate monotone many-to-many alignments between string pairs with one of the procedures outlined in Sections 3 and 4. Then, we train a linear chain conditional random field (CRF; see (Lafferty et al., 2001)) as a graphical model for string transduction on the aligned data. The choice of CRFs is arbitrary; any transduction procedure $\mathsf{tr}$ would do, but we decide for CRFs because they generally have good generalization properties. In all cases, we use window sizes of three or four to predict $\mathbf{y}$ string elements from $\mathbf{x}$ string elements.

## 6.2 Segmentation

Our overall decoding procedure is as follows. Given an input string $\mathbf{x}$, we exhaustively generate all possible segmentations of $\mathbf{x}$, feeding the segmented strings to the CRF for transduction and evaluate each individual resulting sequence of 'graphones' with an $n$-gram model learned on the aligned data, taking the $\mathbf{y}$ string corresponding to the graphone sequence with maximal probability as the most likely transduced string for $\mathbf{x}$. We illustrate in Algorithm 3.

**Algorithm 3** Decoding

1: **procedure** DECODE($\mathbf{x} = x_1 \dots x_m$; $k^*$, $a$, $b$, $\mathsf{tr}$)
2: $Z \leftarrow \emptyset$
3: **for** $s \in \mathcal{C}(m, k^*, a, b)$ **do** ▷ $\mathcal{C}(m, k^*, a, b)$ : the set of all integer compositions of $m$ with $k^*$ parts, each between $a$ and $b$
4: $\hat{\mathbf{y}} \leftarrow \mathsf{tr}(s)$
5: $z_{\hat{\mathbf{y}}} \leftarrow \text{ngramScore}(\mathbf{x}, \hat{\mathbf{y}})$
6: $Z \leftarrow Z \cup \{z_{\hat{\mathbf{y}}}\}$
7: **end for**
8: $z_{\hat{\mathbf{y}}^*} \leftarrow \max_{z_{\hat{\mathbf{y}}}} Z$
9: **return** $\hat{\mathbf{y}}^*$
10: **end procedure**

As to the size of the search space that this procedure entails, note that any segmentation of a string $\mathbf{x}$ of length $n$ with $k$ parts uniquely corresponds to an *integer composition* (a way of writing $n$ as a sum of non-negative integers) of the integer $n$ with $k$ parts, as illustrated below,

| | | ph | | oe | | n | | i | | x |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | = | 2 | + | 2 | + | 1 | + | 1 | + | 1 |

It is a simple exercise to show that there are $\binom{n-1}{k-1}$ integer compositions of $n$ with $k$ parts, where by $\binom{n}{k}$ we denote the respective binomial coefficient. Furthermore, if we put restrictions on the maximal size of parts — e.g. in G2P a reasonable upper bound $l$ on the size of parts would probably be 4 — we have that there are $\binom{k}{n-k}_l$ integer compositions of $n$ with $k$ parts, each between 1 and $l$, where by $\binom{k}{n}_{l+1}$ we denote the respective *$l$-nomial* or *polynomial coefficient* (Comtet, 1974). To avoid having to enumerate segmentations for all possible numbers $k$ of segment parts of a given input string $\mathbf{x}$ of length $n$ — these would range between 1 and $n$, entailing $\sum_{k=1}^{n}\binom{n-1}{k-1} = 2^{n-1}$ possible segmentations in total in the case without upper bound[1] — we additionally train a 'number of parts' prediction model with which to estimate $k$; we call this in short *predictor model*.

To illustrate the number of possible segmentations with a concrete example, if $\mathbf{x}$ has length $n = 15$, a rather large string size given the values in Table 1, there are

$$2472, 2598, 1902, 990, 364, 91, 14, 1$$

possible segmentations of $\mathbf{x}$ with $k = 8, 9, 10, 11, 12, 13, 14, 15$ parts, each between 1 and 4.

## 7 Experiments

We conduct our experiments on three G2P data sets, the German Celex (G-Celex) and French Brulex data set (F-Brulex) taken from the Pascal challenge (van den Bosch et al., 2006), and the English Celex dataset (E-Celex). Furthermore, we apply our algorithms to the four German morphology data sets discussed in Dreyer et al. (2008), which we refer to, in accordance with the named authors, as rP, 2PKE, 13SIA and 2PIE, respectively. Both for the G2P and the morphology data, we hold monotonicity, by and large, a legitimate assumption so that our approach would appear justified. As to the morphology data sets, we illustrate in Table 2 a few string pair relationships that they contain, as indicated by Dreyer et al. (2008).

[1] In the case of upper bounds, Malandro (2012) provides asymptotics for the number of restricted integer compositions, which are beyond the scope of the present work, however.

| | |
|---|---|
| 2PKE. | abbreche**t**, entgegentrete**t**, zuzieh**t** |
| z. | ab**zu**brechen, entgegen**zu**treten, zu**zu**ziehen |
| rP. | redet, r**eibt**, tr**eibt**, verbind**t** |
| pA. | **ge**redet, **ge**r**ieben**, **ge**tr**ieben**, verb**u**nd**en** |

Table 2: String pairs in morphology data sets 2PKE and rP (omitting 2PIE and 13SIA for space reasons) discussed by (Dreyer et al., 2008). Changes from one form to the other are in bold (information not given in training). Adapted from Dreyer et al. (2008).

| | |
|---|---|
| E-Celex | $\{(1,1),(2,1),(3,1),(4,1),(1,2)\}$ |
| rP | $\{(0,2),(1,1),(1,2),(2,1),(2,2)\}$ |
| 2PKE | $\{(0,2),(1,1),(2,1),(2,2)\}$ |
| 13SIA | $\{(1,1),(1,2),(2,1),(2,2)\}$ |
| 2PIE | $\{(1,1),(1,2)\}$ |

Table 3: Data set and choice of $S$. Note that for all three G2P data sets, we select the same $S$, exemplarily shown for E-Celex. The choice of $S$ for rP and 2PKE is taken from Eger (2012). For 13SIA and 2PIE we use comprehensive $\Omega$'s with largest step $(2,2)$ but the algorithm ends up using just the outlined set of steps.

### 7.1 Alignments

We generate alignments for our data sets using Algorithms 1 and 2 and, as a comparison, we implement an exhaustive search bigram scoring model as indicated in Section 4 in an EM-like fashion similar as in Algorithm 2, employing the CMU SLM toolkit (Clarkson and Rosenfeld, 1997) with Witten-Bell smoothing as $n$-gram model. For Algorithm 1, which we also refer to as unigram model in the following, we choose steps $S$ as shown in Table 3. As similarity measure $\mathsf{sim}$, we use $\log \mathsf{prob}$ with Good-Turing smoothing and for $\mathsf{q}$ we likewise use $\log \mathsf{prob}$; we outline the choice of $\mathsf{L}$ below. Initially, we set $\gamma$ and $\chi$ to zero. As an alignment quality measure we consider conditional entropy $H(L \mid P)$ (or $H(P \mid L)$) as suggested by Pervouchine et al. (2009). Conditional entropy measures the average uncertainty of a (grapheme) substring $L$ given a (phoneme) substring $P$; apparently, the smaller $H(L \mid P)$ the better is the alignment because it

| | Perplexity | $H(L \mid P)$ |
|---|---|---|
| 2PKE-Uni | $7.002 \pm 0.04$ | $\mathbf{0.094 \pm 0.001}$ |
| 2PKE-Bi | $\mathbf{6.865 \pm 0.02}$ | $0.141 \pm 0.003$ |
| rP-Uni | $9.848 \pm 0.09$ | $\mathbf{0.092 \pm 0.003}$ |
| rP-Bi | $\mathbf{9.796 \pm 0.05}$ | $0.107 \pm 0.006$ |
| Brulex-Uni | $22.488 \pm 0.35$ | $\mathbf{0.706 \pm 0.002}$ |
| Brulex-Bi | $\mathbf{22.215 \pm 0.21}$ | $0.725 \pm 0.003$ |

Table 4: Conditional entropy vs. $n$-gram perplexity ($n = 2$) of alignments for different data sets. In bold: Statistically best results. $K = 300$ throughout.

produces more consistent associations.

In the following, all results are averages over several runs, 5 in the case of the unigram model and 2 in the case of the bigram model. Both for the bigram model and the unigram model, we select $K$, where $K \in \{50, 100, 300, 500\}$, training samples randomly in each EM iteration for alignment and from which to update probability estimates.

In Figure 2, we show learning curves over EM iterations in the case of the unigram and bigram models, and over training set sizes. We see that performance, as measured by conditional entropy, increases over iterations both for the bigram model and the unigram model (in Figure 2), but apparently alignment quality decreases again when too large training set sizes $K$ are considered in the case of the bigram model (omitted for space reasons); similar outcomes have been observed when similarity measures other than $\log$ prob are employed in Algorithm 1 for the unigram model, e.g. the $\chi^2$ similarity measure (Eger, 2012). To explain this, we hypothesize that the bigram model (and likewise for specific similarity measures) is more susceptible to overfitting when it is trained on too large training sets so that it is more reluctant to escape 'non-optimal' local minima. We also see that, apparently, the unigram model performs frequently better than the bigram model.

The latter results may be partly misleading, however. Conditional entropy, the way Pervouchine et al. (2009) have specified it, is a 'unigram' assessment model itself and may therefore be incapable of accounting for certain 'contextual' phenomena. For example, in the 2PKE and rP data, we find position dependent alignments of the following type,

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| – | g | e | b | t | | g | e | – | b | t |
| ge | g | e | b | en | | g | e | ge | b | en |

where we list the linguistically 'correct', due to the prefixal character of *ge* in German, alignment on the left and the 'incorrect' alignment on the right. By its specification, Algorithm 1 *must* assign both these alignments the same score and can hence not distinguish between them; *the same holds true for the conditional entropy measure*. To address this issue, we evaluate alignments by a second method as follows. From the aligned data, we extract a random sample of size 1000 and train an $n$-gram graphone model (that can account for 'positional associations') on the residual, assessing its perplexity on the held-out set of size 1000. Results are shown in Table 4. We see that, in agreement with our visual impression at least for the morphology data, the alignments produced by the bigram model seem to be slightly more consistent in that they reduce perplexity of the $n$-gram graphone model, whereas conditional entropy proclaims the opposite ranking.

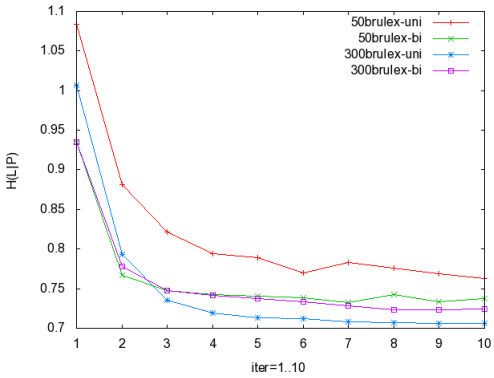


Figure 2: Learning curves over iterations for F-Brulex data, $K = 50$ and $K = 300$, for unigram and bigram models.

### 7.1.1 Quality q of steps

In Table 5 we report results when experimenting with the coefficient $\gamma$ of the quality of steps measure q. Overall, we do not find that increasing $\gamma$ would generally lead to a performance increase, as measured by e.g. $H(L \mid P)$. On the contrary, when choosing as set of steps a compre-

hensive $\Omega$ as in Table 5, where we choose $\Omega = \{(a,b) \mid a \leq 4, b \leq 4\} \backslash \{(0,0)\}$, for $\gamma = 0$, we find values of $0.278$, $0.546$, $0.662$ for $H(L \mid P)$ for G-Celex, F-Brulex and E-Celex, respectively, while corresponding values for $\gamma = 10$ are $0.351$, $0.833$, $1.401$. Contrarily, $H(P \mid L)$, the putatively more indicative measure for transduction from $\mathbf{x}$ to $\mathbf{y}$, has $0.499$, $0.417$, $0.598$ for $\gamma = 0$ and $0.378$, $0.401$, $1.113$ for $\gamma = 10$, so that, except for the E-Celex data, $\gamma = 10$ apparently leads to improved $H(P \mid L)$ values in this situation, while $\gamma = 0$ seems to lead to better $H(L \mid P)$ values.

In any case, from a model complexity perspective,[2] increasing $\gamma$ may certainly be beneficial. For example, Table 5 shows that with $\gamma = 0$, Algorithm 1 will select up to $15$ different steps for the given choice $\Omega$, most of which seem linguistically questionable. On the contrary, with a large $\gamma$, Algorithm 1 employs only four resp. five different steps for the G2P data; most importantly, among these are $(1,1)$, $(2,1)$ and $(3,1)$, all of which are in accordance with linguistic reasoning as e.g. outlined in Section 5.

#### 7.1.2 Colors

We shortly discuss here a possibility to detect latent classes via the concept of colored paths. Assume that a corpus of colored alignments is available and let each color be represented by the contexts (graphones to the left and right) of its members; moreover, define the 'likelihood' $\mathsf{L}$ that the pair $p_{x,y} := (x_{\alpha_{i-1}+1}^{\alpha_i}, y_{\beta_{i-1}+1}^{\beta_i})$ is of color $c$ as the (document) similarity (in an information retrieval sense) of $p_{x,y}$'s contexts with color $c$, which we can e.g. implement via the cosine similarity of the context vectors associated with $p_{x,y}$ and $c$. For number of colors $C = 2$, we then find, under this specification, the following kinds of alignments when running Algorithms 1 and 2 with $\gamma = 0$ and $\chi = 1$,

| a | nn | u | al | | ph | o | n | e | me |
|---|---|---|---|---|---|---|---|---|---|
| & | n | jU | l | | f | @U | n | i | m |
| 1 | 0 | 1 | 1 | | 0 | 1 | 0 | 1 | 0 |

where we arbitrarily denote colors by $0$ and $1$, and use original E-Celex notation for phonemic characters. It is clear that the algorithm has detected some kind of consonant/vowel distinction on a phonemic level here. We find similar kinds of latent classes for the other G2P data sets, and for the morphology data, the algorithm learns (less interestingly) to detect word endings and starts, under this specification.

### 7.2 Transductions

We report results of experiments on transducing $\mathbf{x}$ strings to $\mathbf{y}$ strings for the G2P data and the morphology data sets. We exclude E-Celex because training the CRF with our parametrizations (e.g. all features in window size of four) did regularly not terminate, due to the large size of the data set ($>$ 60,000 string pairs). Likewise for computing resources reasons,[3] we do not use ten-fold cross-validation but, as in Jiampojamarn et al. (2008), train on the first $9$ folds given by the Pascal challenge, testing on the last. Moreover, for the G2P data, we use an $\epsilon$-scattering model with steps $S = \{(1,0),(1,1)\}$ as a predictor model from which to infer the number of parts $k^*$ for decoding and then apply Algorithm 3.[4] For alignments, we use in all cases Algorithms 1 and 2. As reference for the G2P data, we give word accuracy rates as announced by Bisani and Ney (2008), Jiampojamarn et al. (2007), and Rama et al. (2009), who gives the Moses 'baseline' (Koehn et al., 2007).

For the morphology data we use exactly the same training/test data splits as in Dreyer et al. (2008). Moreover, because Dreyer et al. (2008) report all results in terms of window sizes of 3, we do likewise for this data. For decoding we do not use a (complex) predictor model here but rely on simple statistics; e.g. we find that for the class 13SIA, $k^*$ is always in $\{m-2, m-1, m\}$, where $m$ is the length of $\mathbf{x}$, so we apply Algorithm 3 three times and select the best scoring $\hat{\mathbf{y}}$ string. To avoid zeros in the decoding process (see discussion in Section 6.2), we replace the $(0,2)$ steps used in the rP and 2PKE data sets by a step $(1,3)$.

Results are shown in Table 6. Note that, for the G2P data, our approach always outperforms the

[2] Taking into account model complexity is, for example, in accordance with Occam's razor or Akaike's information criterion.

[3] E.g. a single run of the CRF on the G-Celex data takes longer than 24 hours on a standard PC.

[4] We train the $\epsilon$-scattering model on data where all multi-character phonemes such as *ks* are merged to a single character, as obtained from the alignments as given by Algorithms 1 and 2.

| | (1,1) | (2,1) | (3,1) | (4,1) | (1,2) | (1,0) | (2,3) | (3,2) | (3,3) | (4,2) | (4,3) | (4,4) | (2,2) | (0,1) | (1,3) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G-Celex | 86.50 | 11.61 | 1.77 | - | 0.10 | - | - | - | - | - | - | - | - | - | - |
| | 86.14 | 8.17 | 1.63 | 0.02 | 0.00 | 2.56 | 0.10 | 0.04 | 0.01 | 0.09 | 0.91 | 0.28 | - | - | - |
| F-Brulex | 78.85 | 15.08 | 5.85 | - | - | - | 0.20 | - | - | - | - | - | - | - | - |
| | 75.64 | 13.80 | 2.52 | 0.36 | 0.07 | 5.07 | 0.29 | 0.10 | 0.02 | 0.38 | 1.01 | 0.68 | - | - | - |
| E-Celex | 88.87 | 6.58 | 3.05 | - | - | - | - | - | - | - | - | - | - | 1.29 | 0.18 |
| | 75.54 | 8.45 | 0.75 | 0.04 | 1.48 | 4.57 | 0.41 | 0.03 | 0.16 | 0.44 | 2.03 | 3.03 | 0.00 | 2.87 | 0.12 |

Table 5: Steps and their frequency masses in percent for different data sets for $\gamma = 10$ (top rows) and $\gamma = 0$ (bottom rows), averaged over two runs. We include only steps whose average occurrence exceeds 10.

best reported results for pipeline approaches (see below), while we are significantly below the results reported by Dreyer et al. (2008) for the morphology data in two out of four cases. On the contrary, when 'pure' alignments are taken into consideration — note that Dreyer et al. (2008) learn very complex latent classes with which to enrich alignments — our results are considerably better throughout. In almost all cases, we significantly beat the Moses 'baseline'.

## 8 Conclusion

We have presented a simple and general framework for generating monotone many-to-many alignments that competes with Jiampojamarn et al. (2007)'s alignment procedure. Moreover, we have discussed crucial independence assumptions and, thus, limitations of this algorithm and shown that exhaustive enumeration (among other methods) can overcome these problems — in particular, due to the relatively small search space — in the field of monotone alignments. Additionally, we have discussed problems of standard alignment quality measures such as conditional entropy and have suggested an alternative decoding procedure for string transduction that addresses the limitations of the procedures suggested by Jiampojamarn et al. (2007) and Jiampojamarn et al. (2008).

## References


H. Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723.

H. Baayen, R. Piepenbrock, and L. Gulikers. 1996. *The CELEX2 lexical database. LDC96L14.*

T. Baldwin and H. Tanaka. 1999. *Automated Japanese grapheme-phoneme alignment*. In Proc. of the International Conference on Cognitive Science, 349–354.

A.W. Black, K. Lenzo, and V. Pagel. 1998. *Issues in Building General Letter to Sound Rules*. In The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis. ISCA.

M. Bisani, and H. Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5): 434–451.

P.F. Brown, J. Cocke, J., S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2): 79–85.

P.R. Clarkson and R. Rosenfeld. 1997. *Statistical Language Modeling Using the CMU-Cambridge Toolkit*. Proceedings ESCA Eurospeech.

L. Comtet. 1974. *Advanced Combinatorics*. D. Reidel Publishing Company.

A. Content, P. Mousty, and M. Radeau. 1990. Une base de données lexicales informatisée pour le francais écrit et parlé. *L'Année Psychologique*, 551-566.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38.

M. Dreyer, J. Smith, and J. Eisner. 2008. *Latent-Variable Modeling of String Transductions With Finite-State Methods*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Honolulu, Hawaii.

S. Eger. 2012. Sequence alignment with arbitrary steps and further generalizations, with applications to alignments in linguistics. Submitted.

L. Galescu, and J.F. Allen. 2001. *Bi-directional Conversion between Graphemes and Phonemes using a joint n-gram model*. Proc. 4th ISCA Tutorial and ResearchWorkshop on Speech Synthesis. Perthshire, Scotland.

S. Heubach and T. Mansour. 2004. Compositions of $n$ with parts in a set. *Congressus Numerantium*, 164: 127–143.

H.H. Hoang, S.N. Kim, and M.-Y. Kan. 2009. *A Re-examination of Lexical Association Measures*. MWE '09 Proceedings of the Workshop on Mul-

| | CRF-3 | CRF-4 | CRF-4* | DSE-F | DSE-FL | Moses3 | Moses15 | M-M+HMM | BN | MeR+A* |
|---|---|---|---|---|---|---|---|---|---|---|
| F-Brulex | | 93.7 | **94.6** | | | | | 90.9 | 93.7 | 86.7 |
| G-Celex | | 91.1 | **92.6** | | | | | 89.8 | | 90.2 |
| 2PKE | 79.8 | 80.9 | | 74.7 | **87.4** | 67.1 | 82.8 | | | |
| rP | 74.1 | 77.2 | | 69.9 | **84.9** | 67.6 | 70.8 | | | |
| 13SIA | 85.6 | 86.5 | | 82.8 | **87.5** | 73.9 | 85.3 | | | |
| 2PIE | **94.6** | 94.2 | | 88.7 | 93.4 | 92.0 | 94.0 | | | |

Table 6: Data sets and word accuracy rates in percent. **DSE-F**: Dreyer et al. (2008) using 'pure' alignments and features. **DSE-FL**: Dreyer et al. (2008) using alignments, features and latent classes. **Moses3, Moses15**: Moses system with window sizes of 3 and 15, respectively, as reported by Dreyer et al. (2008). **M-M+HMM**: Many-to-many aligner with HMM and instance-based segmenter for decoding as reported by Jiampojamarn et al. (2007). **BN**: Bisani and Ney (2008) using a machine translation motivated approach to many-to-many alignments. **MeR+A***: Results of Moses system on G2P data as reported by Rama et al. (2009). **CRF-3** Our approach with window size of 3 and 3-gram ngram scoring model (see Algorithm 3). **CRF-4**: Our approach with window size of 4 and 3-gram scoring model. **CRF-4***: Our approach with window size of 4 and 4-gram scoring model and 2-best lists. In bold: Best results (no statistical tests). Underlined: best results using 'pure' alignments.


tiword Expressions: Identification, Interpretation, Disambiguation and Applications.

S. Jiampojamarn, G. Kondrak, and T. Sherif. 2007. *Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion*. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007), Rochester, NY, April 2007, 372–379.

S. Jiampojamarn, C. Cherry and G. Kondrak. 2008. *Joint Processing and Discriminative Training for Letter-to-Phoneme Conversion.* 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT), 905–913, Columbus, OH, June 2008.

S. Jiampojamarn, and G. Kondrak. 2010. *Letter-Phoneme Alignment: An Exploration.* The 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pp. 780–788, Uppsala, Sweden. July 2010.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. Proc. 18th International Conf. on Machine Learning, 282–289.

M.E. Malandro. 2012. Asymptotics for restricted integer compositions. Preprint available at http://arxiv.org/pdf/1108.0337v1.

M. Mohri. 2002. Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3): 321–350.

S.B. Needleman and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48: 443–453.

V. Pervouchine, H. Li, and B. Lin. 2009. *Transliteration alignment*. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapora. Association for Computational Linguistics, 136–144.

T. Rama, A.S. Kumar, and S. Kolachina. 2009. *Modeling Letter to Phoneme Conversion as a Phrase Based Statistical Machine Translation Problem with Minimum Error Rate Training*. NAACL HLT 2009 Student Research Workshop, Colorado, USA.

P. Taylor. 2005. *Hidden Markov Models for grapheme to phoneme conversion*. Proceedings of the 9th European Conference on Speech Communication and Technology 2005.

J.D. Updyke. 2008. A unified approach to algorithms generating unrestricted and restricted integer compositions and integer partitions. *Journal of Mathematical Modelling and Algorithms.*

A. van den Bosch, S.F. Chen, W. Daelemans, R.I. Damper, R.I., K. Gustafson, Y. Marchand, and F. Yvon. 2006. *Pascal letter-to-phoneme conversion challenge*. http://www.pascalnetwork.org/Challenges/PRONALSYL.

G. Wimmer, R. Köhler, R. Grotjahn, and G. Altmann. 1994. Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1:98–106.

W. Zucchini. 2000. An introduction to model selection. *Journal of Mathematical Psychology*, 44:41–61.