Proceedings of The Institute of Acoustics

Alignment of Phonemes with their Corresponding Orthography

S.G.C.Lawrence and G.Kaye

Speech Research, IBM UK Scientific Centre,
Athelstan House, St Clement Street, WINCHESTER, SO23 9DR

Introduction

The original motivation for developing this algorithm arose during the automated production of a computer-based dictionary of phonemic transcription based on the *Collins English Dictionary* [1], using a copy of the photo-typesetting tape kindly supplied by Wm. Collins and Co.. Collins's dictionary gives the phonemic transcription of all headwords, and follows the usual lexicographic conventions for representing inflected and derived forms within a headword entry, i.e. by giving partial spellings, and in some cases the pronunciation of the inflectional or derivational suffix. Where the inflected or derived form is considered to be regular no pronunciation is given and the user must apply rules given in the introduction to the dictionary to determine the pronunciation. The alignment algorithm has been used to combine the dictionaries pronunciation of a root with the pronunciation of the suffix (also given in the dictionary) to produce the pronunciation of the inflected or derived form, and for checking all pronunciations so derived. It is a mechanism for combining partial pronunciations and not a means of producing a phonemic transcription *ab initio*. Typical examples of the use of the algorithm in building the dictionary entries are:-

1. The entry for **go** is

   go (/gəʊ/) *vb* **goes**, **go+ing**, **went**, **gone**

   For our purposes pronunciations for **goes** and **going** are derived by applying the rules given in the introduction to the dictionary. The plausibility of the pronunciation so produced is checked by applying the algorithm to the result. If the pronunciation and the orthography cannot be aligned the pronunciation is considered invalid and must be supplied manually. The irregular forms **went** and **gone** appear as separate head words and so do not need to have their pronunciations derived.

2. The algorithm can be used to create complete phonemic transcriptions from partial transcriptions. For example the dictionary entry

   aba+cus /æbəkəs/ pl. **-ci** /-saɪ/

   The '+' denotes a syllable boundary, and in this case also marks the place at which **cus** is to be replaced by **ci**. There is no such marker in the phonemic transcription. It is necessary to know where the syllable boundary is in the transcription so that /kəs/ can be replaced by /saɪ/. The algorithm is used for this. After alignment **aba** is found to correspond with /æbə/ and **cus** with /kəs/. The syllable boundary is between /æbə/ and /kəs/. The end of the orthography can be changed to **ci** and the end of the phonemic transcription to /saɪ/ thus producing the generated form **abaci** /æbəsaɪ/.

3. Where a phonemic transcription is given in *Collins English Dictionary* the lexical stress is indicated as part of it. Words without phonemic transcriptions have stress marked in the orthography. It is therefore necessary to transfer the stress marking to the generated phonemic transcription. The algorithm is used to check the plausibility of the generated phonemic transcription and also to determine the placement of the stress markers. Thus for the word **sug'gester** the phonemic transcription /sədʒεstə/ was generated by appending /ə/ to the transcription /sədʒεst/ of **suggest**. The pronunciation and orthography of **sug'gest** are aligned by the algorithm as:

```
       s   u   g'g e   s   t
   /   s   ə   dʒ  ε   s   t     /
```

this can be used to infer that the stress should be placed on /dʒ/ (using the convention that when the lexical stress in the orthography is between two gemminate consonants, it is placed before the corresponding consonant in the phonemic transcription), giving the derived dictionary entry **suggester** /sə'dʒεstə/.

4. Where a dictionary entry contains alternative orthographies for a head word, but only one phonemic transcription, the algorithm is used to decide if the transcription is plausible for all orthographies. In *Collins English Dictionary* are the following entries.

   abridgement /əbrɪdʒmənt/ or **abridgment**

   abutment /əbʌtmənt/ or **abuttal**

   The pronunciation of **abridgement** and **abridgment** are the same. However, **abuttal** is not pronounced /əbʌtmənt/, nor does any English word contain **al** pronounced as /mənt/. The algorithm indicates that the phonemic transcription of **abutment** is not plausible for **abuttal**, and so its pronunciation must be supplied manually.

The Algorithm

The alignment algorithm is driven by a table of RP phonemes and their corresponding orthography. A different table is required for each accent of English. The present one was deduced with the aid of *Walker's Rhyming Dictionary of the English Language* [2], and Collins English Dictionary. The following is a small section of the table and gives example words. The complete table is given later (sec "The Table of Correspondences").

/-/	⇔	e	have
/ε/	⇔	e	when
/ɪ/	⇔	e	before
/i:/	⇔	e	be
/ə/	⇔	e	government
/eɪ/	⇔	e	crochet
/εə/	⇔	e	whereas
/i·/	⇔	e	recipe
/ɒ/	⇔	e+n	gendarme

The symbol ⇔ denotes *corresponds with.*

When a correspondence occurs only in a specific following orthographic context, this is shown separated from the grapheme by a + (e.g. **e** corresponds to /ɒ/ only in the context of a following **n**). The grapheme following the + must be separately aligned.

Where a grapheme can be considered to have no pronunciation then the symbol /-/ is used to represent the *silent phoneme*. The following multi-graphic phonemes are treated as a single unit and are each stored internally as a single code point:- ɪəʊ, əʊə, iː, iˑ, ɑː, ɔː, uː, ɜː, eɪ, aɪ, ɔɪ, əʊ, aʊ, ɛə.

To improve the robustness of the algorithm, an attempt has been made to minimise the number of *silent* graphemes. There are five cases in the table, and all have high frequencies of occurrence in the LOB corpus [3]. The most common of these graphemes is **e**, occurring largely in a morph or word-final position. Except for these five, all other potentially *silent* graphemes have been combined with one or more adjacent graphemes.

For example:

/m/	⇔	**mb**	**climb**
/s/	⇔	**ps**	**psychology**
/w/	⇔	**wh**	**which**
/f/	⇔	**ph**	**phone**
/ə/	⇔	**eur**	**amateur**
/ʊə/	⇔	**ur**	**sure**
/ɔː/	⇔	**ar**	**war**
/ɜː/	⇔	**or**	**work**

Such combinations are usually well founded for one of the following reasons.

1. They follow a long established orthographic convention e.g., **th** ⇔ /θ,ð/, **ch** ⇔ /tʃ/, **sh** ⇔ /ʃ/.

2. They are phonetically motivated based on other accents of English. One example is **wh** ⇔ /w/ which in Scottish accents is realised as [ʍ] and may be represented phonemically as /hw/ or /ʍ/. Another example is post vocalic r which in RP usually results in lengthening of the vowel, whilst in rhotic accents, such as American English, produces rhoticization of the vowel. It was decided to append the **r** to the preceding vowel when the **r** does not precede a vowel phoneme as in **award** /əwɔːd/ hence a silent **r** cannot occur after a vowel. However when the **r** is intervocalic as in **boring** /bɔːrɪŋ/, the **r** corresponds to /r/.

Those grapheme clusters which cannot be justified for the above reasons are either idiosyncrasies of English spelling, or a consequence of a foreign loan word. They usually have a low frequency of occurrence.

The alignment of **-tion** presents some difficulties. As a matter of principle the alignment **t** ⇔ /ʃ/, **i** ⇔ /-/, **o** ⇔ /ə/, **n** ⇔ /n/ was excluded, leaving two other possibilities. The alignment chosen is **t** ⇔ /ʃ/, **io** ⇔ /ə/ instead of **ti** ⇔ /ʃ/, **o** ⇔ /ə/. For whilst the consonant alternation in word sequences such as **permit, permission, permissive**, is governed by the /t/ in the underlying phonemic form, it is less consistent for the purposes of this algorithm, to group **ti** rather than **io**, for word sequences such as the following:

```
  p er m i t                     d i r e c t
/ p ə  m ɪ t  /                / d ɪ r ɛ k t /
  p er m i ss i o n              d i r e c t i o n
/ p ə  m ɪ ʃ  ə   n /          / d ɪ r ɛ k ʃ ə   n /
  p er m i ss i v e              d i r e c t i v e
/ p ə  m ɪ s  ɪ v - /          / d ɪ r ɛ k t ɪ v - /
```

The algorithm matches the word and its phonemic transcription against the table of correspondences. This table is stored in the computer such that the longest grapheme clusters are scanned first. The initial objective is to match the largest number of graphemes that produces the corresponding phonemes. When a match is found a space is inserted after the last of the matched graphemes and also their corresponding phonemes. If a match is not found then a scan of the graphemes which correspond to the silent phoneme is performed. If a correspondence is found then a '-' is inserted into the phonemic transcription string. If no correspondence is found then there is either an error in the table or in the phonemic transcription and therefore no alignment is possible. If the correspondence consists of multiple graphemes or multiple phonemes then a note is kept of the position in the word and phonemic transcription where the cluster occurs. If, during the left to right scan of the word and transcription, no correspondence can be found and there has been a previous correspondence with more than one grapheme or phoneme then the length of the matched graphemic or phonemic cluster is decreased by one, and a search is made of the correspondence table with the length of the matching string limited to the length of the shortened cluster.

If the cluster is reduced to a single grapheme or phoneme and no match found, then there is either an error in the correspondence table or the phonemic transcription.

The following example illustrates the algorithm. For **create** ⇔ /kriːeɪt/, the first attempt at aligning the orthography with the corresponding phonemes produces

/k/	⇔	**c**	
/r/	⇔	**r**	
/iː/	⇔	**ea**	eg. **each**
/-/	⇔	**t**	eg. **often**
/eɪ/	⇔	**e**	eg. **crochet**

with

/t/ unaligned

The **ea** cluster is split and a second attempt to match the graphemes with the phonemes is attempted starting with the **e** that follows the r. Although /eɪ/ is a cluster that could be split, it is not because the RP phoneme /e/ occurs only in the diphthong /eɪ/. The splitting of a cluster into non-RP phonemes is prevented by using the input conversion table described above that causes such clusters to be treated as single entities.

The second attempt produces

/k/	⇔	c	
/r/	⇔	r	
/i:/	⇔	e	eg. **be.**
/eɪ/	⇔	a	eg. **made.**
/t/	⇔	t	
/-/	⇔	e	eg. **have.**

which gives

```
  c   r   e   a   t   e
/ k   r   i:  eɪ  t   - /
```

Testing the Algorithm

The algorithm has been tested against 32,049 words of a phonemically tagged version of the *Lancaster-Oslo/Bergen Corpus* (see "The Table of Correspondences").

31,773 words were aligned without error.

276 words could not be aligned because they fell into one of the following categories.

- Words containing diacritics eg. **attaché**. This could be avoided if the appropriate correspondence were to be added to the table.
- Words given in Collins with transcriptions containing non-RP phonemes eg. **Truffaut** /tryfo/. It was decided to keep a table of only RP phoneme-to-grapheme correspondences.
- Partial transcriptions. eg. **Sunbury-on-Thames** /sʌnbəri·/. It seems odd that the dictionary entry does not give the pronunciation of **Thames** as well as that of **Sunbury**.

Alignment Examples

The following examples illustrate the performance of the algorithm.

```
 a d v e n t  ur e
/ə d v ɛ n tʃ ə  - /
 a d v i  s e
/ə d v aɪ z - /
 a ff e c t io n
/ə f  ɛ k ʃ ə  n /
 a n g l i c a n
/æ ŋ g l ɪ k ə n /
 a s l ee p
/ə s l i: p /
 a war d
/ə wɔ: d /
 b ar e
/b ɛə - /
```

```
 b or s t a l
/b ɔ: s t ə l /
 c a me r a
/k æ mə r ə /
 c o n v e n t io n
/k ə n v ɛ n ʃ ə  n /
 c y  c l e
/s aɪ k l - /
 d i r e c t io n s
/d ɪ r ɛ k ʃ ə  n z /
 d ir t y
/d ɜ: t i· /
 d i sh
/d ɪ ʃ  /
```

```
 e ss e x
/ɛ s  ɪ ks /
 e  v e
/i: v - /
 e x  a mi  n e
/ɪ gz æ mɪ  n - /
 f o n d
/f ɒ n d /
 h i th er t o
/h ɪ ð  ə  t u: /
 h o r i  z on
/h ə r aɪ z n̩  /
 h u rr y
/h ʌ r  i· /
```

```
 i  d e n t i c al
/aɪ d ɛ n t ɪ k l̩  /
 l e ss on
/l ɛ s  n̩  /
 mea t
/mi: t /
 n o  t a b l e
/n əʊ t ə b l̩ - /
 n ou n
/n aʊ n /
 o b s er v er
/ə b z ɜ: v ə  /
 p e n al t y
/p ɛ n l̩  t i· /

 p l a t f or m
/p l æ t f ɔ: m/
 p or t u  gu e  s e
/p ɔ: t jʊ g  i: z - /
 p o ss e ss io n
/p ə z  ɛ ʃ  ə  n/
 p r ai s e
/p r eɪ z - /
 p r e s t i  g e
/p r ɛ s t i: ʒ - /
 s e tt l e me n t
/s ɛ t  l̩ - m ə n t /
 s p ar e
/s p ɛə - /

 s ur r e n d er
/s ə  r ɛ n d ə  /
 u l t i ma t e l y
/ʌ l t ɪ mɪ t - l i· /
 v a  gu e
/v eɪ g  - /
 v i s i t or
/v ɪ z ɪ t ə  /
 v o  c al
/v əʊ k l̩  /
 wer e n ' t
/w ɜ: - n - t /
 wor th y
/w ɜ: ð  i· /
```

The Table of Correspondences

The table that follows shows the phoneme-to-grapheme correspondences together with some statistics gathered by using the algorithm on a lexicon. Only part of the table is given. The complete table will be published at a later date. The lexicon used is those 32.049 words in the *LOB Corpus* that are not homographs and whose pronunciation has been derived from *Collins English Dictionary*. The frequencies of the words are taken from the *LOB Corpus*. The table is in the order in which it is scanned when matching the orthography with the phonemic transcriptions.

The **Phone** column shows the phonemes that correspond to the graphemes in the column entitled **Graph**. Characters other than IPA symbols indicate the following :

- - The graphemes in the **Graph** column do not correspond with a phoneme (i.e. are silent).
- * If the grapheme **r** follows the graphemes in the **Graph** column in a word and the phoneme */r/* does not follow the phonemes in the **Phone** column in the transcription then the grapheme **r** forms part of the correspondence.

In the **Graph** column, characters other than graphemes are used to indicate a condition which has to be satisfied for there to be a correspondence. The characters and their meaning are as follows :

- + The graphemes following the ' + ' must follow the graphemes that precede the ' + ' for the correspondence to occur.
- > The graphemes following the ' > ' must not follow the graphemes that precede the ' > ' for the correspondence to occur.
- - The graphemes following the '-' must precede the graphemes that precede the '-' for the correspondence to occur.
- < The graphemes following the ' < ' must not precede the graphemes that precede the ' < ' for the correspondence to occur.

- May be substituted for one of the characters . , / ' or a word end.

The **Example** column gives the most frequently occurring word in the lexicon in which the correspondence occurs. The **# Text** column gives the cumulative sum of the frequencies of occurrence of those words in the corpus (LOB), from which the lexicon was produced, in which the correspondence occurs. If the correspondence does not occur in the lexicon (i.e. 0 in this column), it is listed because it occurs in words with RP phonemic transcriptions in *Collins English Dictionary*. The **# Lex** column gives the number of times the correspondence occurred in the lexicon.

Phone	Graph	Example	# Text	# Lex
-	t>ed	often	579	32
-	s+.	cannes	0	0
-	r-r	referred	130	8
-	h	john	1748	174
-	e>r	have	139867	7594
aɪə*	y	byron	65	8
aɪə*	oi	choir	10	3
aɪə	iro+n	iron	79	7
aɪə*	ie	society	800	45
aɪə*	ia	trial	546	86
aɪə*	i+r	required	1627	126
aɪ	y	by	10824	331
aɪ	uy	buy	148	8
aɪ	igh	might	4335	181
aɪ	ia	vial	0	0
aɪ	i	i	34269	2096
aɪ	ey	eyes	550	12
aɪ	eigh	height	83	4
aɪ	ei	either	544	40
aɪ	ay	ramayana	11	5
aɪ	ai	cairo	38	17
aɪ	ae	maestro	2	1
aɪ	a	bandaranaike	2	1
aʊə*	ow	dowry	1	1
aʊə*	ou	our	1654	18
aʊə*	au	gaur	0	0
aʊ	ow	now	5940	184
aʊ	ough	plough	25	7
aʊ	ou	out	10918	464
aʊ	eo	macleod	49	2
aʊ	aw	haworth	1	1
aʊ	au	strauss	29	11
aʊ	aou	caoutchouc	0	0
aʊ	ao	laos	26	3
æ	i	meringue	2	2
æ	ai	laing's	4	4
æ	ach	drachm	0	0
æ	aa	naali	0	0
æ	a'a	ma'am	5	1
æ	a	and	94264	4539
ɑːr	r+.	r.m.s.	17	9
ɑː	r+.	r.	169	11
ɑː	ia	acciaccatura	0	0
ɑː*	ea	heart	303	25

Phone	Graph	Example	# Text	# Lex
ɑː	e+m	contretemps	1	1
ɑː	e+n	genre	4	1
ɑː*	e	sergeant	58	10
ɑː	au	laughed	213	17
ɑː	ard	boulevards	1	1
ɑː	al	half	483	32
ɑː	agh	omagh	0	0
ɑː*	ae	stranraer	6	2
ɑː*	aa	afrikaans	15	7
ɑː*	a	are	24710	1548
biː	b+.	b	378	22
b	pb	cupboard	22	4
b	bb	abbey	410	144
b	b	by	54975	4016
diː	d+.	d	191	25
dʌbljuː	w+.	w.	96	10
d	ld	could	0	0
d	dh	oldham	14	2
d	dd	added	1674	154
d	d	and	153411	8977
d	bd	bdellium	0	0
dʒeɪ	j+.	j.	145	8
dʒiː	g+.	g.	207	20
dʒ	jj	hajji	0	0
dʒ	j	just	4912	405
dʒ	gi	religious	378	36
dʒ	gg	suggested	402	15
dʒ	g	general	11034	1028
dʒ	dj	adjustment	129	27
dʒ	dg	knowledge	1057	123
dʒ	d	soldiers	197	11
dʒ	ch	norwich	23	5
eɪtʃ	h+.	h.	161	13
eɪ	ez	laissez	0	0
eɪ	eyo	eyot	0	0
eɪ	ey	they	3900	44
eɪ	es	demesnes	1	1
eɪ	eigh	eight	444	38
eɪ	ei	reign	125	30
eɪ	ee	beethoven	27	7
eɪ	ed	passepied	0	0
eɪ	eagh	castlereagh	0	0
eɪ	ea	great	1169	21
eɪ*	e	crochet	200	56

Phone	Graph	Example	# Text	# Lex	Phone	Graph	Example	# Text	# Lex
eɪ	ay	may	8574	284	eɪ	aigh	straight	134	7
eɪ	au	gauge	22	4	eɪ	ai	main	6515	481
eɪ	ao	gaol	17	3	eɪ	ae	phaeton	8	5
eɪ	ais	renaissance	17	5	eɪ	a	made	33492	2930

Summary

The aligned forms of words and their phonemic transcriptions, together with statistics on frequency of rule use, are seen as invaluable aids in developing a set of text-to-phoneme rules, where the ordering of rules, based on frequency of occurrence is critical. The aligned forms may also be useful in speech recognition systems.

The table of correspondences may be used in the development of aids for the teaching of English spelling. From the table can be deduced the most likely orthography for a given sound.

References

[1] 'Collins English Dictionary', William Collins, Glasgow, (1979).

[2] J. Walker, 'The Rhyming Dictionary of the English Language', Routledge & Keegan Paul, London, (1924).

[3] 'LOB (Lancaster-Oslo/Bergen) Corpus'. Department of English, University of Oslo, Oslo, (1978).

Acknowledgements

It is a pleasure to thank William Collins Sons and Co. for making available for our use a copy of the *Collins English Dictionary* in computer-readable form.