

## Towards an automated classification of Englishes

Søren Wichmann and Matthias Urban

### 1. Introduction<sup>1</sup>

As first steps towards rethinking a topic of research it is fruitful to divorce one's approach from the conventional ones, to develop new methodologies, and to apply them in consistent ways ignoring divisions of fields of research imposed by the tradition. In this paper we apply a relatively new methodology, that of an automated lexicostatistics using the so-called "Levenshtein distance", to varieties of English, treating dialects and creoles in essentially the same way. Thus, we follow the example of Kortmann and Schneider (2004) and some of the other contributors to this handbook (cf. chapters by A. Schneider, by E. Schneider, and by Winford) in including both Englishes traditionally understood to be "dialects of English" and Englishes understood to be creoles or pidgins (henceforth "creoles"). We do treat the two categories distinctly, using a finer phonetic resolution for the dialects than for the creole data, but the basic methodology is the same. Using the same set of data, we then go on to look at the distribution of diversity among dialects of English. Studying phonetic diversity in global varieties of English is relevant for the history of the language, and for rethinking this history.

---

<sup>1</sup> We are grateful to Eric W. Holman and Viveka Velupillai for computational help and help with the data, respectively, as well as to Peter Trudgill for correcting us in some of the dialectological matters. The data used for the creole classification are found in Wichmann et al. (2010b).

An important next step in the rethinking process is to integrate results with existing knowledge. Our results for English dialectology consolidate observations in the literature arrived at by quite different methods. With regard to creoles, the literature does not offer any principled classification. We hope to show that the same approach that we apply to dialects can help to remedy this, even if the creole classification that we will present is not strictly speaking a phylogenetic one. While the data used are synchronic we show how the classification by computational methods mirrors well known facts of the history of English.

## 2. Methodology

A simple method for comparing languages or language varieties has been developed within the Automated Similarity Judgment Program (ASJP) project. The method involves the comparison of words with the same meaning and the measurement of phonological differences between such words using a version of the so-called Levenshtein distance (LD). The LD measures the number of substitutions, deletions, and insertions needed to turn one string of symbols into another one. Subsequently, distances between pairs of speech varieties are established by averaging distances among the different words compared. This approach was introduced by dialectologists working on Irish (Kessler 1995) and Dutch (Heeringa 2004), and has more recently been applied to several other languages and language families. Work devoted to experiments with refinements of the LD that take into account degrees of phonetic difference between the phonemes in the strings compared has shown that sophisticated weighting schemes do little towards

increasing the accuracy of the ensuing classifications (Heeringa 2004: 186; Heeringa et al. 2006).<sup>2</sup>

More specifically, our approach is the following. We use fine-grained Unicode IPA transcriptions as our starting point, where the data from variants of English are the 110-item word lists of Heggarty, Maguire, and McMahon (2007), which are available online. The words are then transformed into X-SAMPA (Wells 1995) using an automated transcription system (developed by Henrik Theiling, <http://www.theiling.de/ipa/>). We then calculate an average normalized Levenshtein distance (as in Serva and Petroni 2008) for the set of words compared. Finally the resulting matrix of distances among dialects is subjected to the NeighborNet algorithm (as implemented in SplitsTree 4, Huson and Bryant 2006). This ‘pipeline’ of arriving at a classification is described in more detail in the Online Materials, which are intended to make our results replicable. The attention to fine phonetic detail is undoubtedly necessary for studying the configuration of dialects, and a substantial number of words used in the comparisons is likely also an asset. For producing phylogenies of less closely related language variants, however, it has been shown (Holman et al. 2008) that using just 40 historically stable form-concept pairings produces near-optimal results. Thus, for the comparison of creoles we have used the 40-

---

<sup>2</sup> A method for comparing words which is quite different from the LD has been developed by Heggarty, McMahon, and McMahon (2005) and McMahon et al. (2007). This method requires cognate segments to be aligned with reconstructed proto-segments, and only then is a distance metric applied to the cognate segments. A full description of the distance metric is not available, however, so results such as those of Heggarty, McGuire, and McMahon (2010) cannot be replicated, but at least they can be compared with those of our own method (Section 3 below).



Figure 1. A NeighborNet of English dialects based on data in Heggarty, Maguire, and McMahon (2007) (AAVE = African American Vernacular English; EMGT = emergent; Trad = Traditional; Typ = typical)

The brackets in Figure 1 have been added to visualize phylogenetic clusters which also represent major geographical clusters. The geographical pattern (visualizing the same clusters) is shown in Figure 2.

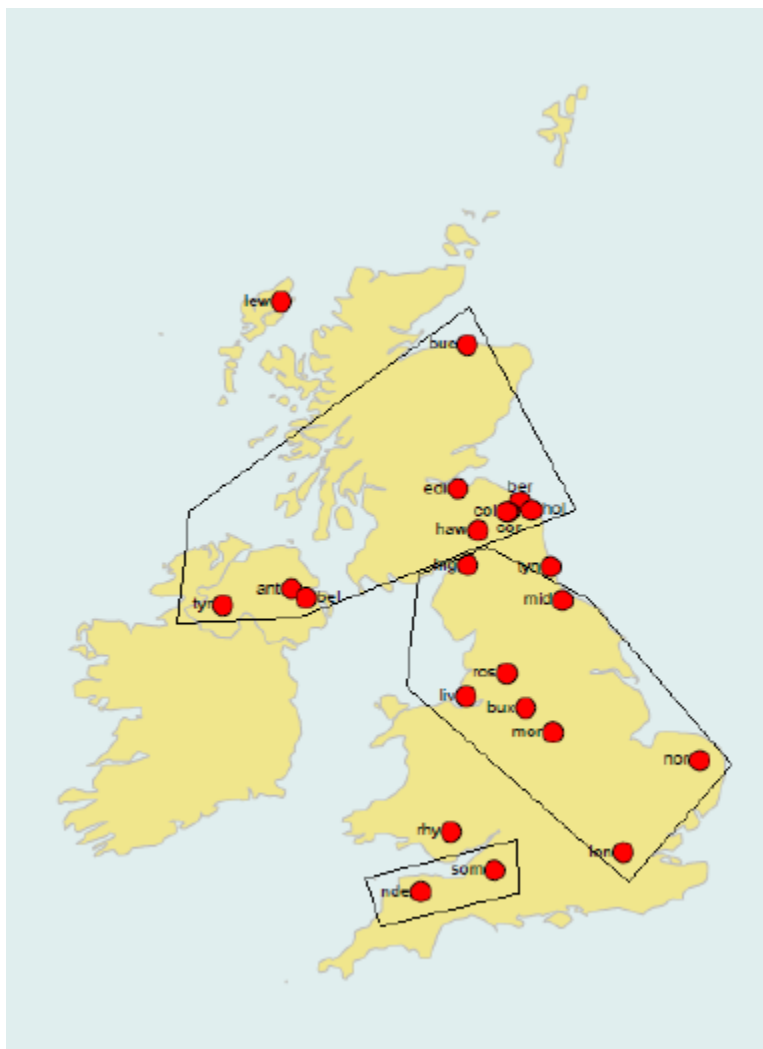


Figure 2. Locations of dialects covered in this study, with indications of major splits<sup>3</sup>

We obtain a basic split between the dialects of Scotland and England, which is also well established by dialectologists (e.g., Aitken 1992; Trudgill 1999). This is due mainly to the more conservative nature of Scots (the most notorious isogloss is the retention of the velar fricative *x* in many English varieties of Scotland). Douglas (2010: 43) also notes the stronger legacy of Old Norse in Scotland as compared with England.

Particularly noteworthy in our results is the fact that dialects spoken in the immediate vicinity of the modern geographical border, but on the English side, i.e., Berwick, Cornhill and Holy Island, cluster together with the Scottish dialects rather than with the English ones. This contrasts with the classification proposed in Heggarty, Maguire, and McMahon (2010), which suggests a patterning of the dialects entirely in line with modern geographical borders. Our result, however, is what one would expect given the history of the Scottish-English borderland. For instance, the history of Berwick-upon-Tweed, situated 5 km south of the Scottish-English border, is particularly checkered. Its political affiliation changed frequently between Scotland and England before it was finally annexed by England in the fifteenth century (Llamas 2010: 230). Llamas, Watt, and Johnson (2009: 387) note that “Berwick English ... is clearly a hybrid of Scottish and Northumbrian varieties, as would be expected given the high levels of historical and contemporary contact between the populations on either side of the border.”

---

<sup>3</sup> The abbreviations of the localities, which are spelled out fully in Figure 1, are self-explanatory.

At the next classificatory level Trudgill (1999: 66) calls for a basic subdivision of the Southern dialects into those of the Southwest, said to form a “recognizable unity”, and those of the East. One of the arguments is the rhoticity of the western dialects as opposed to the lack of it in the East. The unity of the Southwestern dialects is robustly mirrored in our results.

We note that Ireland patterns with Scotland and Northern England in Figure 1. This is in line with aspects of Irish history. In order to secure English control over the island, several controlled waves of settlement by English people and, in particular in Ulster, by Scots, were instigated by the English government from the sixteenth century onward. Rather than being a continuation of earlier Middle Irish, Modern Irish English is a product of the spread of the varieties of English spoken by these immigrants, with typological differences between the varieties mirroring the different origins of the settlers (Algeo 2010: 200). Bliss (1977) offers a similar account, while Hickey (2001) maintains that some older features persist, at least on the east coast of Ireland. The historical facts also explain the close ties between dialects of Northern Ireland and Scotland as mirrored in their similar vowel systems. The same vowel system, notably, is also found in northernmost Northumberland, but not in any other dialect of English (Wells 1982: 183). This is in line with our result whereby the English dialects close to the Scottish border cluster with those of Scotland.

Moving beyond the British Isles, we enter the “second diaspora” in the four diaspora system outlined by Kachru and Smith (2008: 5) to describe the historical spread of English around the globe, that is, its spread to North America, Australia, New Zealand, and South Africa. We note that varieties of North American and Australian English

respectively cluster together in separate regions of the network. This behavior reflects differences in the times of departure of emigrants from Britain and the corresponding stages in the history of English. In both varieties the influence from original local languages was minimal (Detering 2010: 385). South Africa is different in this regard. Here there has been a considerable influence from speakers of Afrikaans and from local languages. In this respect South African English is more similar to the varieties issuing from the “third diaspora” of English. This took place during the era of colonization, which saw the spread of English to areas in South and Southeast Asia, Africa, and the Caribbean (and during this time most English-based pidgins and creoles also emerged, see discussion below). Detering (2010: 386) speaks of “huge differences” between varieties emerging from the third diaspora due to influence of local indigenous languages. This explains why varieties representing the third diaspora do not show particularly tight clustering in the network. As a recent addition, Splitstree 4.11.3 includes an option for quantifying the degree to which a language fits into a network through the so-called delta score of Holland et al. (2002). Briefly described, a higher score corresponds to a lesser overall fit of a taxon into a phylogeny. Tellingly, Johannesburg, Nigeria, Singapore, and New Delhi are among the top nine dialects with respect to their delta scores (which range from 0.40 to 0.44), showing that they are among the dialects which are most involved in conflicting phylogenetic signals. The averages for respectively North America, Australia-New Zealand, and the British Isles are all the same (0.38).

#### 4. Results of classification: creoles

The classification of English-based creoles reveals a strong areal signal that mirrors the differing substrate languages and their phonological and lexical characteristics in different regions of the world. In Figure 3 we again present a NeighborNet, this time for the creoles. We observe one particular cluster being markedly distant from the rest, namely the one consisting of creoles spoken in Suriname. This is likely due to the fact that Suriname creoles are not exclusively English-based, but have experienced superstrate influence early on also from Portuguese and later from Dutch. Additionally, Suriname creoles retain a relatively high number of lexemes of African origin (Good 2009: 921).

Elsewhere, we find creoles with Melanesian substrate grouped together with Australian creoles. Within this group, there is a close configuration of Bislama and Tok Pisin. Holm (2000: 96) describes these languages as “modern branches” of a contact variety of English called Melanesian Island Pidgin, which arose in the nineteenth century. Both are closely affiliated with Torres Strait Creole in the network. After its initial pidginization, this variety is known to have been influenced by Melanesian Island Pidgin (Holm 2000: 96). As the nearest neighbors of this cluster we find varieties of Kriol.

Moving counter-clockwise in the network we next encounter a cluster of African-based creoles, with subdivisions into those which originated in North America (Gullah and Geechee, which cluster with Hawaii Creole English), a somewhat looser clustering of those actually spoken in Africa (Ghanaian Pidgin English, Nigerian Pidgin, Kamtok, Pichi, and Krio), and a tight cluster of the Caribbean creoles (Limonese, Jamaican, Vincentian).

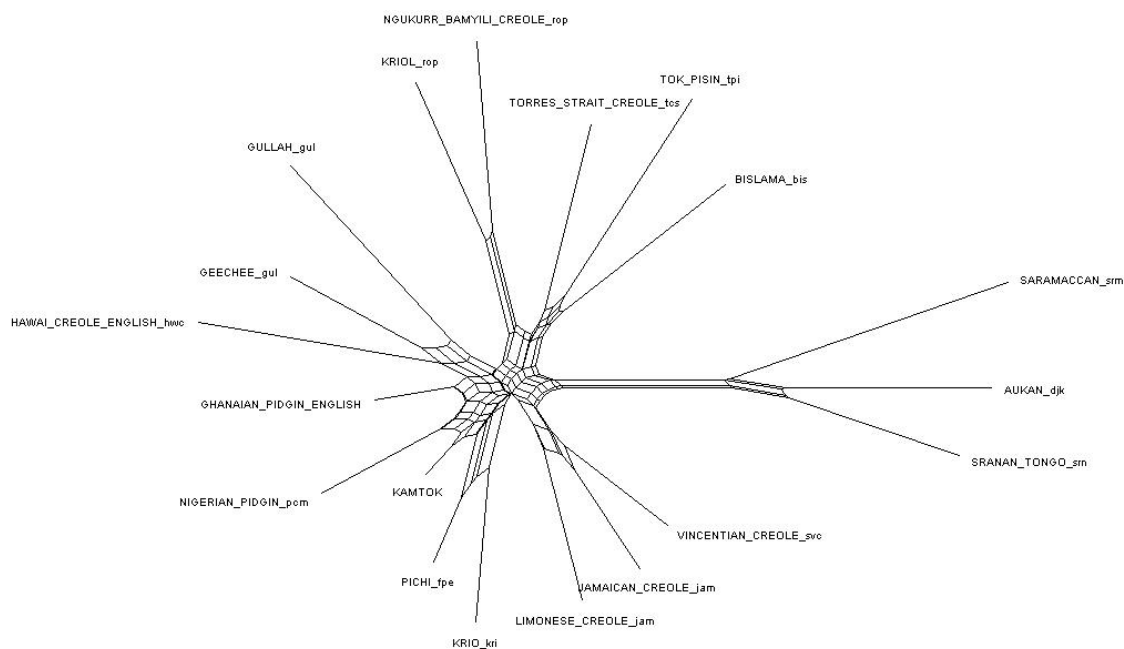


Figure 3. An ASJP classification of some English-based creoles (ISO 639-3 codes are supplied when available; Kamtok is spoken in Cameroon, not far from the coast and the Nigerian border).

It may be questioned whether creoles should be classified genealogically (Thomason and Kaufman 1988). It is nevertheless interesting to carry out the exercise since the network in Figure 3 actually does reveal some historical connections. Moreover, in the larger picture of the world's languages it is hard to deny that English-based creoles are related to English in some sense. In the ASJP World Language Tree of Lexical Similarity (Müller et al. 2010) all the creoles displayed in Figure 3 are gathered under a single node, together with English and Scots.

## 5. Insights from diversity measures

A network such as the one displayed in Figure 1 above is a means of displaying the historical linguistic configuration of a set of related speech varieties. Is it possible, based on the same data, to say something about the geographical trajectories of these language varieties, too? Wichmann, Müller, and Velupillai (2010) suggest an affirmative answer to this question. Based on the common assumption that the region of highest diversity is likely to be at the same time the center of origin of a language group (Sapir 1916), the authors develop a new technique for inferring hypothetical centers of dispersal of groups of related languages. The method consists in assigning diversity indices to each language within a family, where a diversity index is calculated as the mean of the ratios between linguistic and geographical distances holding between the given language and all the others within the group. The linguistic distance is defined as in Section 2 above, and the geographical distance is, for the sake of simplicity, as the crow flies. The center of dispersal of the genealogical linguistic group is identified with the location of the language having the highest diversity index. The results can conveniently be displayed on maps like the one in Figure 4, where shades and sizes of dots represent relative values of diversity indices, dark shades representing the highest, lighter shades lower values, and small dots the lowest. The figure shows that if just the North American variants are used, New York City emerges as the center of diversity.<sup>4</sup>

---

<sup>4</sup> The data that we use include wordlists for “Standard American” and “Standard Canadian”. For the purpose of Figure 4 these are assigned to Los Angeles and Toronto respectively.



Figure 4. Display of differential diversity indices for North American English varieties

When we use the same dataset as was used for inferring the phylogenetic network in Figure 1 the area with the highest diversity turns out to be Edinburgh. The result is the same whether we use all World Englishes or just the varieties confined to the British Isles. Inasmuch as World Englishes did issue from the British Isles the method is on the right track. On the isles themselves, however, the geographical and linguistic resolution appears to be too fine for the results to yield to a meaningful interpretation as a historical signal of dispersal

## 6. Outlook

In addition to developing new methodologies for dealing with traditional problems, such as, in our case, how to classify speech varieties, rethinking may be directed towards new objects of inquiry. Traditional dialectology is concerned with the identification of geographical boundaries and how to distinguish one dialect area from another, while being less concerned with qualitative differences between areas. The methods that produced our Figures 4 (mapping the homeland of North American English) and 1 (the NeighborNet of non-creole varieties of English with its distinctive branch lengths) in

addition trigger the issue of differential diversity. We found that the center of diversity in the specific sense used in this chapter was higher in the British Isles than everywhere else.

Nevertheless, the data that we use only tell part of the story. In a forthcoming paper Bernd Kortmann classifies varieties of English, including creoles, using grammatical features rather than the standard wordlists drawn upon in the present paper. The result obtained is quite a different one, with a markedly greater diversity outside (as opposed to in) the British Isles. In a classification based on such features the varieties on the British Isles all appear on a single twig among the branches of the larger tree.

As far as diversity is concerned, then, the pictures pan out differently depending on the types of data drawn upon. The causes for these differences are too complex to unravel in this contribution, wound up as they are with the different histories and sociolinguistic settings of each variety, but we hope that our observations may inspire future inquiries into the nature of linguistic diversity at the level of closely related speech forms such as those of Englishes throughout the world.

## References

- Aitken, Adam J. 1992. 'Scots'. In *The Oxford Companion to the English Language*, ed. Tom McArthur, 893–99. Oxford: Oxford University Press.
- Algeo, John. 2010. *The Origins and Development of the English Language*. 6th edn. Boston: Wadsworth.
- Bliss, Alan. 1977. 'The Emergence of Modern English Dialects in Ireland'. In *The English Language in Ireland*, ed. Diarmaid O'Muirthe, 7–19. Dublin: Mercier.

- Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. 'Automated Classification of the World's Languages: A Description of the Method and Preliminary Results'. *STUF – Language Typology and Universals* 61: 285–308.
- Detering, David. 2010. 'Variations across English: Phonology'. In *The Routledge Handbook of World Englishes*, ed. Andy Kirkpatrick, 385–99. London: Routledge.
- Douglas, Fiona. 2010. 'English in Scotland'. In *The Handbook of World Englishes*, eds. Braj Kachru, Yamuna Kachru, and Cecil L. Nelson, 41–57. Oxford: Blackwell.
- Good, Jeff. 2009. 'Loanwords in Saramaccan, an English-based Creole of Suriname'. In *Loanwords in the World's Languages: A Comparative Handbook*, eds. Martin Haspelmath and Uri Tadmor, 918–43. Berlin: Mouton de Gruyter.
- Heeringa, Wilbert. 2004. *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. Ph.D. dissertation. Rijksuniversiteit Groningen.
- Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. 'Evaluation of String Distance Algorithms for Dialectology'. In *Linguistic Distances, ACL Workshop held at ACL/COLING*, ed. John Nerbonne and Erhard Hinrichs, 51–62. Sydney Shroudsburg, PA: ACL.
- Heggarty, Paul, Warren Maguire, and April McMahon. 2007. *Accents of English from Around the World*. Accessed 17 February 2010.  
<http://www.soundcomparisons.com>.

- 2010. ‘Splits or Waves? Trees or Webs? How Divergence Measures and Network Analysis Can Unravel Language Histories’. *Philosophical Transactions B* 365: 3829–43.
- Heggarty, Paul, April McMahon, and Robert McMahon. 2005. ‘From Phonetic Similarity to Dialect Classification: A Principled Approach’. In *Perspectives on Variation*, ed. Nicole Delbecque, Johan van der Auwera, and Dirk Geeraerts, 43–91. Berlin: Mouton de Gruyter.
- Hickey, Raymond. 2001. ‘The South-East of Ireland. A Neglected Region of Dialect Study’. In *Language Links: The Languages of Scotland and Ireland*, ed. John Kirk and Dónall Ó Baoill, 1–22. Belfast: Queen’s University.
- Holland, B. R., K. T. Huber, A. Dress, and V. Mouton. 2002.  $\delta$  plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution* 19: 2051–2059.
- Holm, John A. 2000. *An Introduction to Pidgins and Creoles*. Cambridge: Cambridge University Press.
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. ‘Explorations in Automated Language Classification’. *Folia Linguistica* 42: 331–54.
- Huson, Daniel H. and David Bryant. 2006. ‘Application of Phylogenetic Networks in Evolutionary Studies’. *Molecular Biology and Evolution* 23: 254–67.
- Kachru, Yamuna and Larry E. Smith. 2008. *Cultures, Contexts, and World Englishes*. New York: Routledge.

- Kessler, Brett. 1995. 'Computational Dialectology in Irish Gaelic'. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics*, 60–67. San Francisco: Morgan Kaufmann Publishers.
- Kortmann, Bernd. Forthcoming. 'How Powerful Is Geography as an Explanatory Factor of Variation? First Evidence from the World Atlas of Variation in English'. In *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives* (working title), ed. Peter Auer, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi. Berlin: Walter de Gruyter.
- Kortmann, Bernd and Edgar Schneider, in collaboration with Kate Burridge, Rajend Mesthrie, and Clive Upton, eds. 2004. *A Handbook of Varieties of English*. Berlin: Mouton de Gruyter.
- Llamas, Carmen. 2010. 'Convergence and Divergence across a National Border'. In *Language and Identities*, ed. Carmen Llamas and Dominic Watt, 227–236. Edinburgh: Edinburgh University Press.
- Llamas, Carmen, Dominic Watt, and Daniel Ezra Johnson. 2009. 'Linguistic Accommodation and Salience of National Identity in a Border Town'. *Journal of Language and Social Psychology* 28: 381–407.
- McMahon, April, Paul Heggarty, Robert McMahon, and Warren Maguire. 2007. 'The Sound Patterns of Englishes: Representing Phonetic Similarity'. *English Language and Linguistics* 11: 113–42.
- Müller, André, Søren Wichmann, Viveka Velupillai, Cecil H. Brown, Pamela Brown, Sebastian Sauppe, Eric W. Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, Oleg Belyaev, Robert Mailhammer, Matthias Urban, Helen Geyer, and

- Anthony Grant. 2010. 'ASJP World Language Tree of Lexical Similarity: Version 3 (July 2010)'. [http://email.eva.mpg.de/~wichmann/language\\_tree.htm](http://email.eva.mpg.de/~wichmann/language_tree.htm).
- Sapir, Edward. 1916. *Time Perspective in Aboriginal American Culture, a Study in Method*. Ottawa: Government Printing Bureau.
- Serva, Maurizio and Filippo Petroni. 2008. 'Indo-European Languages Tree by Levenshtein Distance'. *EuroPhysics Letters* 81.68005.
- Thomason, Sarah G. and Terrence Kaufman. 1988. *Language Contact, Creolization and Genetic Linguistics*. Berkeley: University of California Press.
- Trudgill, Peter. 1999. *The Dialects of England*. 2nd edn. Oxford: Blackwell.
- Wells, John C. 1982. *Accents of English. An Introduction*. Cambridge: Cambridge University Press.
- 1995. 'Computer-coding the IPA: a proposed extension of SAMPA'. Accessed 6 August 2010. <http://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf>.
- Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010a. 'Evaluating Linguistic Distance Measures'. *Physica A* 389: 3632–39.
- Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, David Beck, and Helen Geyer. 2010b. 'The ASJP Database (version 13)'. <http://email.eva.mpg.de/~wichmann/languages.htm>.
- Wichmann, Søren, André Müller, and Viveka Velupillai. 2010. 'Homelands of the World's Language Families: A Quantitative Approach'. *Diachronica* 27: 247–76.