Digital Scholarship in the Humanities Advance Access published December 8, 2014

Geographical patterns of formality variation in written Standard California English

Costanza Asnaghi, Dirk Speelman and Dirk Geeraerts
Quantitative Lexicology and Variational Linguistics Research Unit,
University of Leuven, Belgium

Abstract

Formality variation in the written use of lexical words in the relational sphere in California English is analyzed on a geographical level for the first time in this article. Linguistic data for word alternations including a formal and an informal term for a specific concept are gathered from newspapers Web sites written in English through site-restricted Web searches across California (Asnaghi, *An Analysis of Regional Lexical Variation in California English Using Site-Restricted Web Searches*. Joint Ph.D. Dissertation, Università; Cattolica del Sacro Cuore and University of Leuven, Milan, Italy and Leuven, Belgium, 2013) and analyzed with a series of spatial statistical analyses (Grieve et al. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, **23**: 193–221, 2011). Urban versus rural and north versus south tendencies are detected in the language choices of California journalists. These tendencies are rooted in the history of the Golden State as well as in its socio-economical structure (Starr and Procter. Americans and the California dream, 1850–1915. *History: Reviews of New Books*, **1**(9): 201–201, 1973; Hayes, *Historical Atlas of California: With Original Maps*. Berkeley/Los Angeles/London: University of California Press, 2007).

Correspondence:
Asnaghi Costanza,
Quantitative Lexicology and
Variational Linguistics
Research Unit,
University of Leuven,
Belgium.
E-mail:
costanza.asnaghi@kuleuven.be

1 Introduction

The geographical distribution on the California territory of a group of linguistic variables formed by variants denoting different degrees of formality in addressing or describing people is surveyed here. The linguistic data under scrutiny were collected from online newspapers from across the state of California. This is the first evaluation to provide a quantitative analysis of regional formality variation in the lexical domain of written Standard California English.

Although California is the most populous state in the USA, previous large-scale dialectology studies have never paid much attention to it. Studies reporting the phonetic situation of California cities or areas have been conducted (e.g. DeCamp, 1971; Hinton et al., 1987; Moonwomon, 1991; Hagiwara, 1995; Eckert, 2000; Waksler, 2000; Bucholtz et al., 2007; Hall-Lew, 2010; Podesva, 2011; Kennedy and Grama, 2012); however, no studies have attempted to give a big picture of language variation in the state. David Reed and Allan Metcalf did attempt to produce a Linguistic Atlas of the Pacific Coast in the 50s (Reed and Metcalf, 1952), aiming at a description of the language in California and Nevada. The 300 interviews conducted for the Atlas demonstrated indeed that California English

doi:10.1093/llc/fqu060

is an independent dialect, as reported by Elizabeth S. Bright (1971); nonetheless, an atlas with the results of these inquiries was never published. This article intends to investigate formality variation in the newspaper register on a large scale in a region that has been linguistically explored only to a limited extent before.

Comprehensive dialect studies of American English have surveyed California as a small portion of the 48 contiguous USA (Carver, 1987; Labov et al., 2006; Grieve, 2009). In particular, Grieve's (2009) analysis was conducted in a quantitative way and included 11 California cities. A further study (Grieve, 2011) referred to the same set of data focusing on the results for contraction, which is a feature of informal writing. Comparisons between Grieve's study and the present research will be discussed below.

Formality variation, i.e. variation in the form of linguistic choices of words in accordance with the conventions of the social context of use, has been extensively acknowledged by various foundational sociolinguistics studies. Labov (1972b) describes the 'principle of formality' as a formal linguistic context obtained whenever a speaker monitors his or her language production. In the context of newspaper writing, language is supposed to be somewhat 'formal' as for Labov's definition. In fact, written language, especially when targeted for publication, is usually monitored by the writer. Douglas Biber (1988) describes linguistic variation as a continuum, and sees formality as a dimension of that continuum. One of the aims of this article is to show that a single dimension is indeed sufficient to pinpoint variation between more formal and less formal linguistic styles.

On the quantitative geographical variational level, Szmrecsanyi (2014) compares quantitative studies based on linguistic atlas data to quantitative studies based on linguistic frequency data (i.e. corpora), concluding that frequency-based approaches provide more realistic linguistic evidence.

Only few previous studies have attempted to analyze formality variation quantitatively in a geographical perspective. Examples are Grieve (2011) and the lexicon-based sociolectometrical approach introduced in Geeraerts et al. (1999) and further developed in Speelman et al. (2003) and Ruette et al. (2011). Grieve (2011) demonstrates that measures of contraction rate, e.g. *would not* and *wouldn't*, are regionally patterned in written Standard American English. The investigation that this article presents mainly differentiates from Grieve's study in that it examines the behavior of lexicon rather than contraction rates. Lexicon is, according to Peter Trudgill (1999), the level of the English language where stylistic differences are most evident.

The remainder of the article is structured as follows: Section 2 presents the employed method of data gathering, Section 3 presents two spatial statistical analyses, and Section 4 provides an overview of the results. Finally, Section 5 reviews the results: an interpretation for the detected regional patterns is provided, and a direction for future developments of this study is also suggested.

2 Linguistic Data

Online newspapers are the selected register for this study, corresponding more generally to the written Standard English register, where the term 'register' is used according to Charles Ferguson's (1994) interpretation, or 'the communicative situation', and the term 'Standard English' follows Trudgill's (1999: 124) definition according to which 'all newspapers that are written in English are written in Standard English'. Online newspapers publish a great amount of freely available text in a computer-readable form and are annotated for place of publication.

While online newspapers texts do not cover all possible registers of written Standard English, online newspaper texts can nonetheless be representative indicators of style for a particular region. In fact, online newspapers texts contain a wide variety of articles and sections, encompassing local, national, and international news, sport reports, arts and culture columns, travel tips, business insights, and in some cases even fiction; they also often include advertisements and readers' opinions. In Biber's (1988) research on register relations, limited to the dimension of ‘involved’ versus ‘informational’ text production that is comparable to the focus of this article, text categories in the newspaper sphere (press reviews, press reportage, and editorials), although very different from a limited category of written texts (i.e. personal letters), share the same formality level with a wide number of texts (i.e. biographies, academic prose, science fiction, religion, humor, and popular lore); newspaper language is also relatively close to further categories of written texts (i.e. official documents and fiction). Moreover, the online newspaper archives that were reached for this research contain text in such a quantity that a distinction among genres is not essential for the determination of patterns of regional lexical variation at a national level of resolution.

The most efficient technique up to date to gather linguistic frequencies from online texts is site-restricted Web searches (Grieve et al., 2013). Starting from a list of suitable newspaper Web sites based in the geographical area to be investigated and from a list of lexical alternation variables formed by variants denoting different degrees of formality, the Google search engine was queried for the number of hits for each variant of the selected variables in the entire archive of each newspaper.

The list of newspaper Web sites included 334 online newspapers based in 273 different California locations (see Asnaghi, 2013 for the full list of newspapers). Daily and weekly California online newspapers written in English with substantial reports on local facts were considered suitable for this study[1]. University, entertainment, and parish papers were excluded. Hits from online newspapers published in the same location were summed.

The list of the lexical alternation variables was formed by 12 nouns denoting people in the family sphere: *dad/father*, *mom/mother*, *grandpa/grandfather*, *grandma/grandmother*, *folks/parents*[2], and *kid/child*. For each variable of this list, the first variant is less formal, and the second variant is more formal.

Other variants for these concepts exist in English. For example, when asked for nicknames for maternal grandmothers, over 10,000 Harvard Dialect Survey respondents provided a varied range of answers, the most frequent being *grandma* (50.67%), while the other variants (*nana*, 5.77%; *grandmother*, 4.78%; *granny*, 3.77%; *grammy/grammie/grammi*, 3.24%; *mimi*, 0.97%; other undisclosed nouns, 30.79%) were infrequent, with a frequency rate lower than 6% in all cases. Those low-frequency variants were not analyzed in this study. In fact, the emphasis of this study is on high-frequency variants. Although some sociolinguists insist that all variables, including low-frequency ones, should be included in a dialect study (Labov, 1972a; Kretzschmar, 2009), in this case, the inclusion or exclusion of low-frequency variants would return a similar proportion in the count of linguistic variation for a specific concept. For example, in Sonoma, California, represented here by the online newspaper *sonomawest.com*, *grandma* occurs 1570 times (74.5%), *grandmother* occurs 503 times (23.8%), *nana* occurs 32 times (1.5%), *grammie* occurs only once (0%), *grammi* and *granny* do not occur at all in the newspaper archive (0%) (*grammy* and *mimi* were omitted from this test because too ambiguous for an effective search through site-restricted Web searches: a search for *grammy* returned hits meaning ‘Grammy Awards’, or the annual award given by the American National Academy of Recording Arts and Sciences for achievement in the record industry; a search for *mimi* returned different proper names). Excluding low-frequency variables does not dramatically change the proportion for the high-frequency ones: in this case, if only the hits for *grandma* and *grandmother* are calculated, *grandma* accounts for 75.7% of the results, while *grandmother* accounts for 24.2% of the results, affecting the percentage only slightly (0.8% change for *grandma*, 0.4% change for *grandmother*).

Moreover, the focus here is on the relational sphere for no specific reason other than accuracy in data collection and interpretation. In fact, *dad/father*, *mom/mother*, *grandpa/grandfather*, *grandma/grandmother*, *folks/parents*, and *kid/child* are relatively unambiguous terms, which is a requirement for good performance in site-restricted Web searches. Other near-synonyms were considered, among which were *nab/arrest*, *shiv/knife*, and *cool/*

excellent. Nonetheless, these word pairs would not perform well in site-restricted Web searches, the words *arrest*, *shiv*, and *cool* being highly polysemous. Furthermore, in word pairs such as *display/manifest*, *quick/rapid*, *stipend/emolument*, *wisecrack/joke*, *determine/ascertain*, and many others that were considered for research, a formality gradient between the first and the second term does exist, but it is not guaranteed that one of the terms is actually used with informal or formal intentions (see Brooke et al., 2010). Therefore, formality variation in the lexicon can be cautiously retrieved only from a vocabulary that corresponds to an aware choice of the writer toward a relatively formal or a relatively informal linguistic variant to identify a specific concept. This is best represented by relational terms, where the alternative between the formal way to address a member of a family, i.e. *father*, and the informal way to address the same person, i.e. *dad*, is well demarcated.

The reason why we opted for site-restricted Web searches over more conventional corpus linguistic techniques to mine newspaper articles is that site-restricted Web searches allow for the examination of a vaster quantity of text. Site-restricted Web searches are queries conducted through a Web search engine that look for a specific term or expression in a specific Web site. Google Search was used in this case, although other Web search engines would also apply. The specific Web search through Google was conducted by entering the tag *site:* immediately followed by the Web site domain (e.g. *sonomanews.com*) and by the target term expressed inside quotation marks (e.g. '*mother*'). Quotation marks prevent from automatic stemming, i.e. a search for pages containing not only the selected term but also closely related variants of the term such as its plural form, etc., and force Google Search to return results that exactly match the searched term. The *www.* prefix was removed from the Web site addresses so that the search engine would search the entire domain, including pages with a different prefix such as *sports.sonomanews.com*. The same search was conducted automatically through a Python script for each term of the six word alternation variables.

Regional linguistic variation in this study is observed in natural language discourse that is produced by a sample of language users (i.e. journalists) taken from the entire population, rather than just by a few long-term residents, as is the case in traditional American dialect surveys. Online newspaper language, restricted to letters to the editor, has been previously analyzed in dialectology (Grieve, 2011). Online newspapers do not usually disclose explicit information on the provenance of the informants. Nonetheless, while it is possible that journalists are not residents of the city where a newspaper is published, this may occur only for a limited number of cases[3], therefore a substantial source of texts will be written by local or near local authors, whereas the rest will be from all over the place, which we consider as noise in the data. In addition, nonlocal journalists writing for a local newspaper produce text that can be either considered as noise for our study, or, more usefully, text produced in and for a local audience, therefore valuable data. For example, the San Francisco Chronicle archive contains a letter to the editor from a newspaper reader in Seattle, WA[4]. Is that letter part of the San Francisco speech community or the Seattle speech community? Probably neither, or both, but the communicative situation for that specific letter is in San Francisco. Therefore, the cases in which journalists are not based in the same location as the newspaper that they write for do not represent a threat to the validity of the site-restricted Web searches method. It should be noted once again that this study aims at measuring regional patterns in the newspaper register of California English rather than at making assumptions on the general speech community of California. Moreover, the quantity of linguistic data collected through this method as well as the statistical techniques used to interpret the data (see Section 3) smooth out the scattered cases of nonresident journalists.

The site-restricted Web searches method may seem deceptive: Web searches will count potential nonequivalent uses of the searched variants. Additionally, search engines return the number of pages in which a target term can be found, rather than the number of actual occurrences of the term in the Web site. Furthermore, Google Search returns only an estimate number of results, and it is hardly possible to define a search boundary according to a specific genre or register other than by selecting specific Web sites—newspaper Web sites in this case.

Despite the potential noise in the collected data, the site-restricted Web searches method was proved valid through an evaluation across the USA for lexical word alternation variables distribution attested by both this method and previous American English research. Site-restricted Web searches returned linguistic distribution results that were comparable to results obtained through traditional linguistic data collections (see details in Grieve et al., 2013). The validity of the method was therefore widely proven despite of the potential noise. It should be noted that other linguistic collections focusing on lexical variation such as Hans Kurath's (1949), E. Bagby Atwood's (1962), and Cassidy and Hall's (Cassidy and Hall, 1985, 1991; Hall and Cassidy, 1996; Hall, 2002, 2012; Hall and von Schneidemesser, 2013) involved noise in dialect data too. Nonetheless, the noise in the data did not undermine the surveys.

Site-restricted Web searches obtain a considerable quantity of linguistic data through automated series of computational instructions as opposed to costly traditional data collection methods. For example, for the variable *grandpa/grandfather*, in the newspaper *Redding Record Searchlight* based in Redding, CA, we found 20,300 hits for the variant *grandpa* and 29,300 hits for the variant *grandfather*. Given the very big amount of data not only on the dimension of examined newspapers but also on the dimension of occurrences for each variant, it would be unrealistic to pursue manual analysis as in traditional corpus linguistics. For any deeper analysis on the textual distribution of the terms, the identification of any collocational patterns, the examination of sub-genre differences, as well as for a thorough cleaning of false hits from the analysis, the application of very refined distributional methods would be required.

3 Statistical Analysis

The frequencies for the six lexical alternation variables were counted through site-restricted Web searches across the 273 locations. The results were then calculated as proportions, providing continuous data for the analysis presented in the sections below.

The computation for the value of every single variable followed Equation 1. In Equation 1, T is the value of the variable to be obtained, N_1 is the value of the first variant of the variable, and N_2 is the value of the second variant of the variable.

$$T = \frac{N_1}{N_1 + N_2}. \qquad (1)$$

The collected data sorted as proportions for each linguistic variable at each location were analyzed through Moran's I, a statistical technique for the measurement of global spatial autocorrelation (Moran, 1948; Odland, 1988; Grieve, 2011).

Moran's I studies phenomena having a random probability distribution in more than one dimension in space. Its foundation is in cross-product statistic (Γ, Equation 2), but it differs from cross-product statistic (Hubert et al., 1981) in that it takes into consideration multiple dimensions. The equation for cross-product statistic is as follows:

$$\Gamma = \sum_i \sum_j W_{ij} C_{ij}, \qquad (2)$$

where i and j are any pair of locations, W_{ij} is the weight between observation i and j, also called spatial weight matrix or neighboring function (Paradis, 2009), and C_{ij} is the measure of the distance between the values of i and j. C_{ij} is calculated according to a certain measure of distance in cross-product statistic (such as Euclidean distance, Manhattan distance, spherical distance, etc.; see Sawada, 2004); in Moran's I, C_{ij} is calculated as displayed in Equation 3, namely as the product of the distance of the value x_i at location i and of the value x_j at location j from the global mean of the z-values.

$$C_{ij} = (x_i - \overline{x})(x_j - \overline{x}) \qquad (3)$$

Also, as for the Pearson statistic, Moran's I includes a scaling factor (expressed here in

Equation 4) that is not present in the cross-product statistic:

$$M = \frac{n}{W \sum (x_i - x)^2}. \quad (4)$$

The complete formula of Moran's *I* is provided in Equation 5 as follows:

$$I = \frac{n \sum \sum W_{ij}(x_i - \overline{x})(x_j - \overline{x})}{W \sum (x_i - x)^2}. \quad (5)$$

Moran's *I* results typically range between −1 and +1 for each variable, where scores toward −1 denote dispersion, scores toward +1 denote clustering, and scores near 0 indicate random distribution. The *p*-values correspond to a one-tailed 0.05 alpha level.

In this study, Moran's *I* measured the level of significance of each lexical alternation variable. In particular, a one-tailed *t* test assessed positive global spatial autocorrelation, establishing whether each variable evince regional clustering.

In Table 1, the scores of Moran's *I* significance test of global spatial autocorrelation, the *z*-scores, and the *p*-values are displayed, ranging from highly significant to less significant. In general, the significance at the global level for all selected variables was considerable.

After the analysis of global spatial autocorrelation, a test of local spatial autocorrelation was conducted. The main difference between global and local spatial autocorrelation statistics is that a global measure of spatial autocorrelation returns a number for each variable of the data set, while a local measure of spatial autocorrelation returns a number associated with each observation unit, as a quantitative expression of Waldo Tobler's (1970: 237) first law of geography: 'Everything is related to everything else, but near things are more related than distant things'.

In order to calculate local spatial autocorrelation, a spatial weighting function has to be defined. A spatial weighting function is a protocol that specifies the weight to the comparison of every pair of locations.

The analyses reported here are based on a reciprocal weighting function, which is a common weighting function that assigns a weight to a comparison based on the reciprocal of the distance between the two locations, so that weight decreases with distance (Odland, 1988).

Table 1 Global spatial autocorrelation results

Alternation	Moran's *I*	z-score	p-value one tail
Mom/mother	0.0713	6.9895	0.0001
Dad/father	0.0629	6.2007	0.0001
Grandpa/grandfather	0.0390	3.9779	0.0001
Folks/parents	0.0375	4.0074	0.0001
Kid/child	0.0298	3.243	0.0005
Grandma/grandmother	0.0237	2.5550	0.0053

The test of local spatial autocorrelation Getis-Ord *Gi* followed Equation 6 (Ord and Getis, 1995; Grieve, 2011):

$$G_i(d) = \frac{\sum_j w_{ij}(d)x_j - W_i\overline{x}(i)}{s(i)\{[((n-1)S_{1i}) - W_i^2]/(n-2)\}^{\frac{1}{2}}}, \quad (6)$$

where $j \neq i$, $S_{1i} = \sum_j w_{ij}^2, (j \neq i), \overline{x}$ and s^2 denote sample mean and variance.

Getis-Ord *Gi* examined each linguistic variable for significant levels of positive or negative local spatial autocorrelation. The goal of this analysis is to determine what distributional values of the variables are found in the surroundings of the chosen locations.

Getis-Ord *Gi* fetched a *z*-score for each variable at each location. Variables returning a *z*-score value larger or equal to ±1.64 were considered locally significant. The *z*-scores were considered significant at a one-tailed 0.05 alpha level.

Getis-Ord *Gi* scores were positively significant or negatively significant for locations surrounded by other locations with similar values or with dissimilar values, respectively. In particular, positive Getis-Ord *Gi* scores indicated that the first lexical variant was relatively more frequent, while negative Getis-Ord *Gi* scores indicated that the second lexical variant was relatively more frequent in that neighborhood. Getis-Ord *Gi* scores approximating to zero indicated a region of variability between a preference for the first lexical variant and a preference for the second lexical variant in that

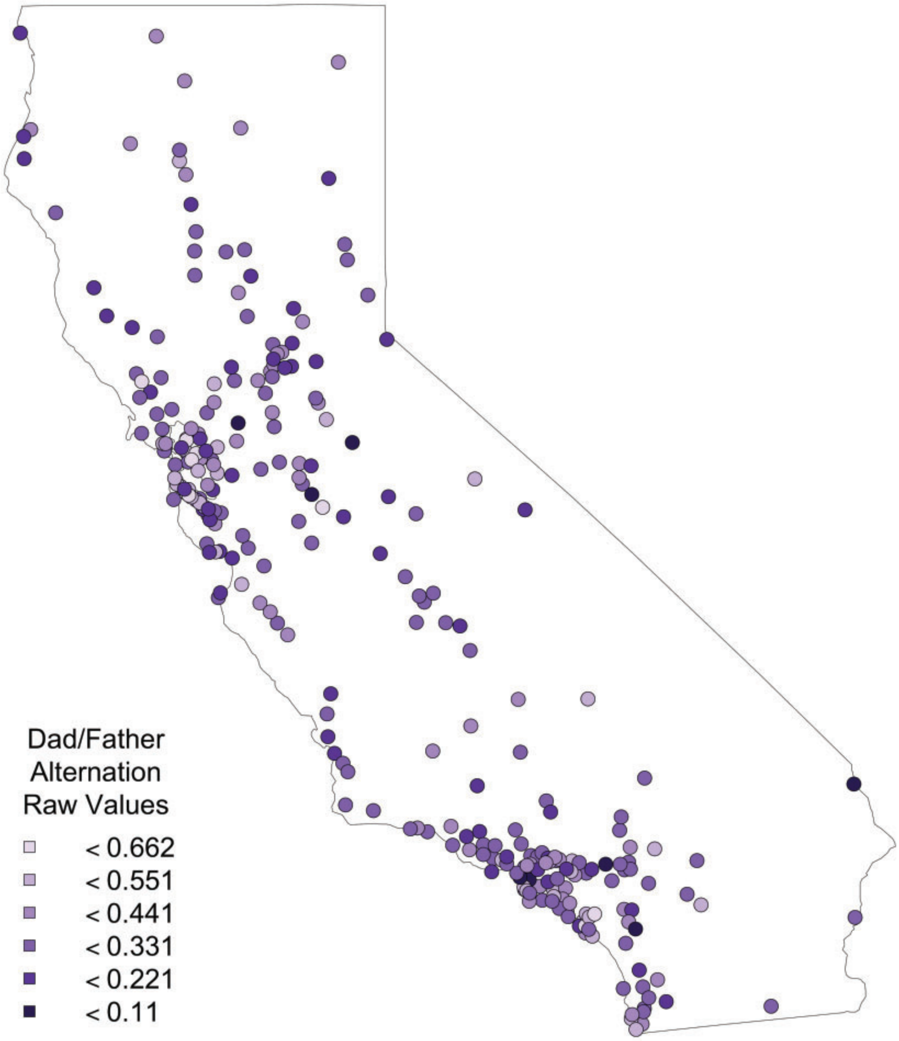

Fig. 1 Probability of *Dad* relative to *Father*. A color version of this figure is available online

neighborhood. For example, Los Angeles City, Hollywood, and Beverly Hills are locations close to each other—the distance among the three cities can be represented by a triangle of sides of, respectively, 7, 4, and 10 miles. For the variable *kid*/*child*, the results of the *Gi* analysis are very similar in the three cities (approximating the results to two decimal places, Los Angeles $Gi = 2.32$, Hollywood $Gi = 2.34$, and Beverly Hills $Gi = 2.39$). The positive results indicate that *child* is relatively more used than *kid* in the newspapers of that area.

Figure 1 is an example of a map of California on which the surveyed locations were plotted in dots filled with shades according to the raw proportion values, while Figure 2 is an example of an autocorrelated map. The map in Figure 1 displays the probability of the first variant relative to the second variant of the continuous lexical alternation variable *dad*/*father* in California English. A dot filled with a lighter color indicates that the first variant is more common in the location identified by that specific dot. A dot filled with a darker color indicates that the second variant is more common in the identified location. The map in

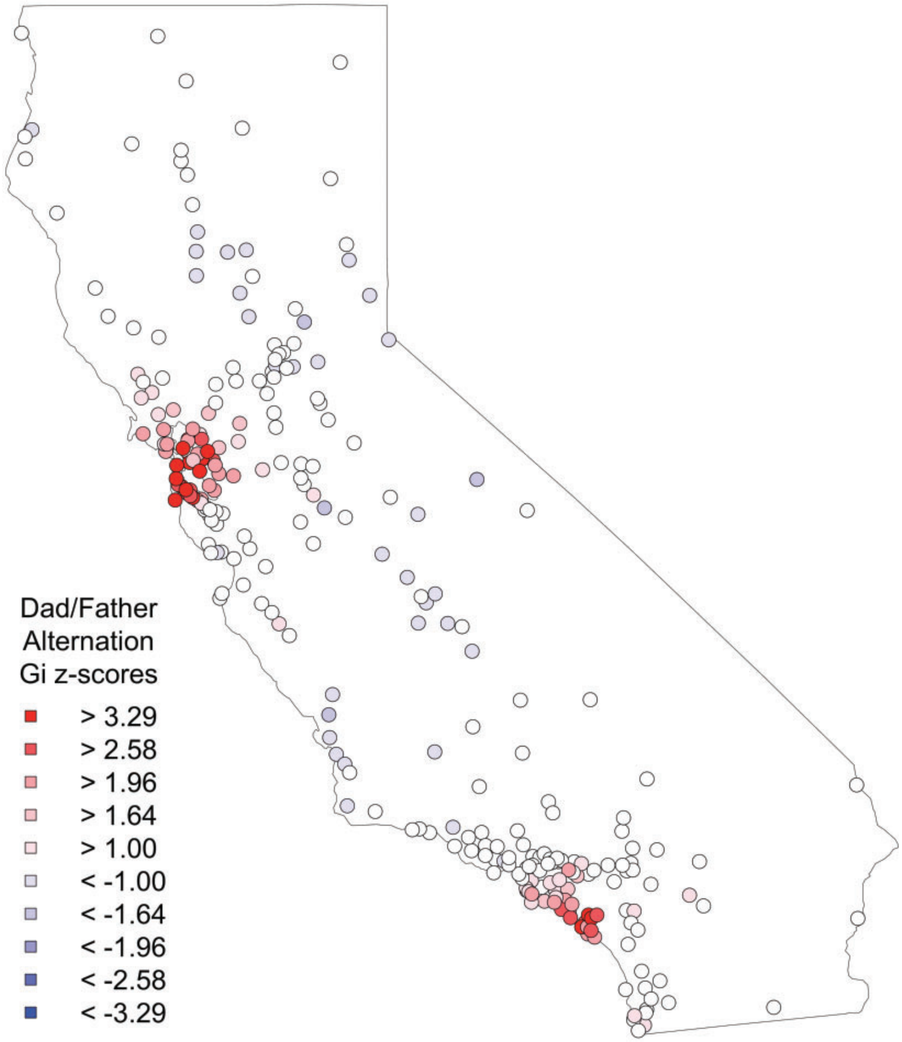

Fig. 2 Probability of *Dad* relative to *Father*: map of local spatial autocorrelated Getis-Ord *Gi z*-score values. A color version of this figure is available online

Figure 2 represents again the probability of the first variant relative to the second variant of the continuous lexical alternation variable *dad/father*, this time by plotting the *Gi z*-scores on the California map. In these maps, a darker dot (or a red dot, with reference to the online color version of the map) indicates that the identified location was associated with a positive *Gi z*-score, and therefore the first variant occurs relatively more frequently in that location. A lighter dot (or a blue dot, with reference to the online color version of the map) proves that the identified location was associated with a negative *Gi z*-score, and therefore the second variant occurs relatively more frequently in that location. A grey dot (or a white dot, with reference to the online color version of the map) shows a region of fluctuation in the preference for the first or second variant. Going back to the example, Los Angeles, Hollywood, and Beverly Hills are represented by darker/red dots in the autocorrelated map for the variable *kid/child* (see Fig. 6) due to the close

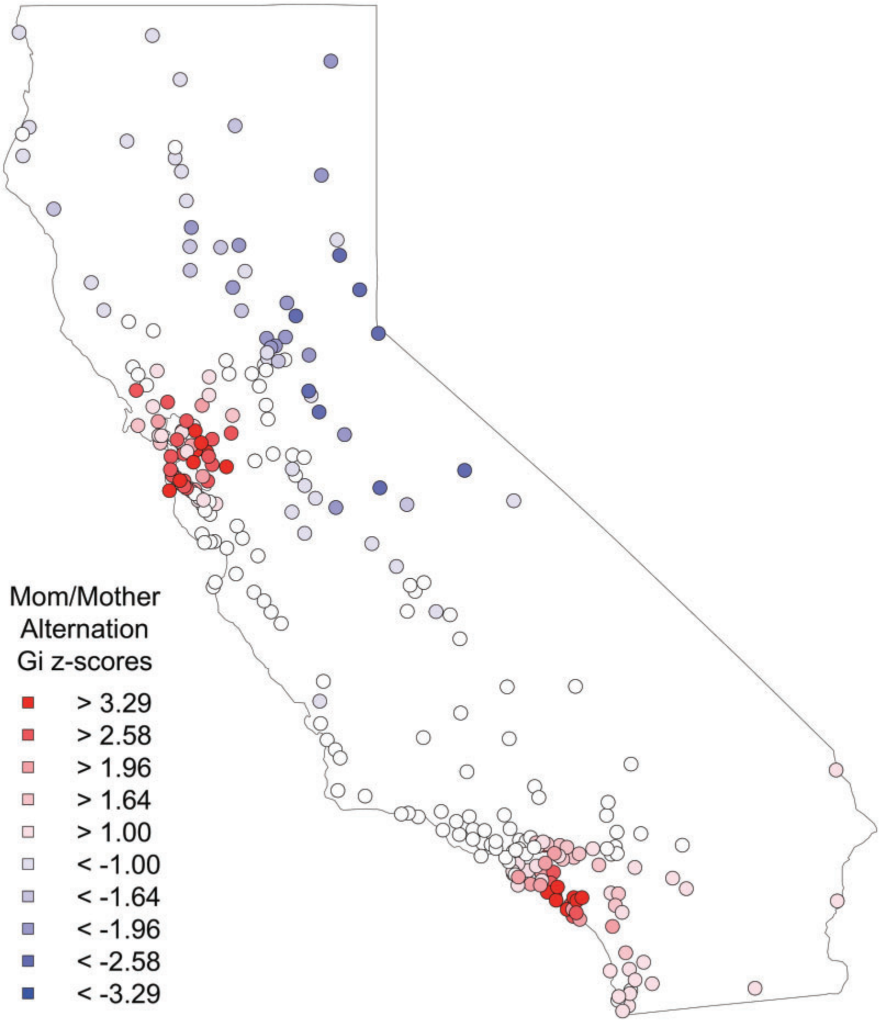

Fig. 3 Probability of *Mom* relative to *Mother*: map of local spatial autocorrelated Getis-Ord *Gi* *z*-score values. A color version of this figure is available online

relatedness of the positive results that the *Gi* analysis returned at those locations.

Figure 2 is visually clearer than Figure 1, and provides an example of the powerful smoothing effect of the employed statistical autocorrelation technique. This analysis is the quantitative equivalent of isoglosses identification (Grieve et al., 2011). In fact, local spatial autocorrelation is a direct way to decipher the linguistic data more clearly, leveling the noise that was present in the raw data.

4 Results

The maps of the Getis-Ord *Gi* *z*-scores (Figs 2–5) exhibit two main tendencies in the language choices of California journalists, namely a north/south and an urban/rural distinction, as described in the following paragraphs. For an overview of the urban/rural areas of California, Figure 7 presents population density information throughout the state.

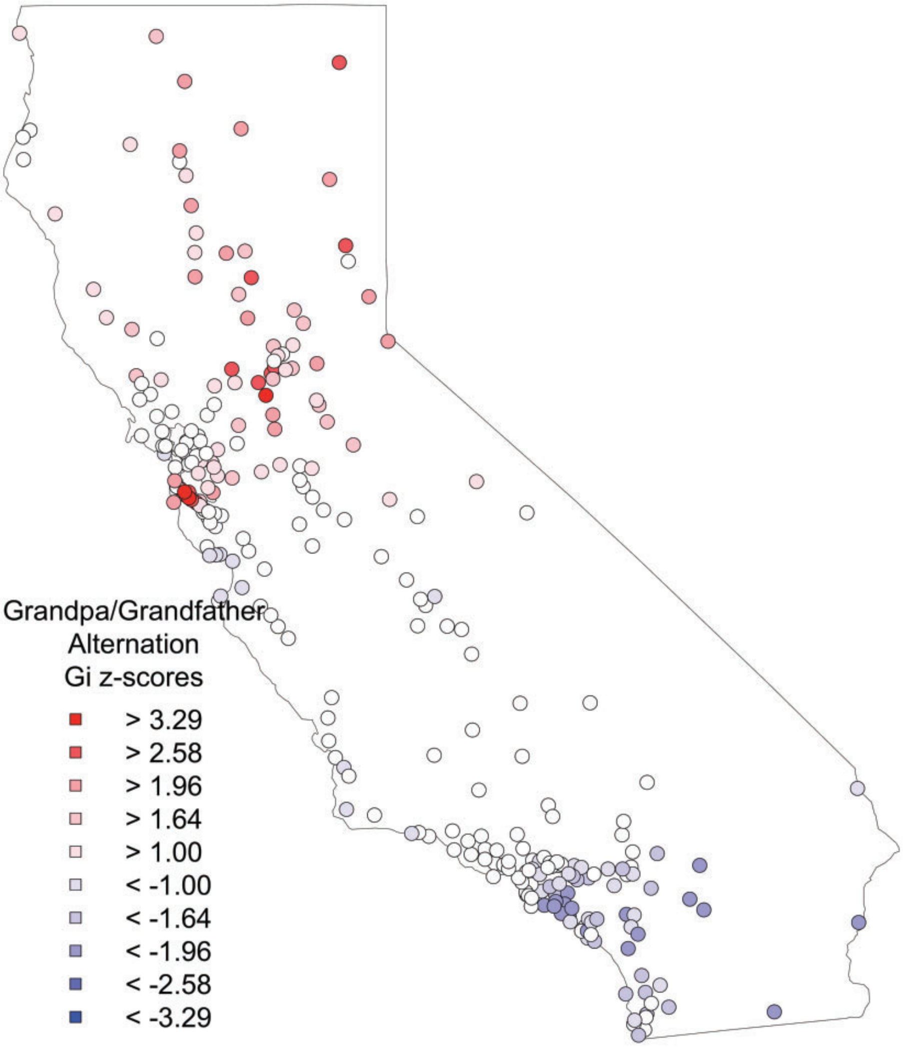

Fig. 4 Probability of *Grandpa* relative to *Grandfather*: map of local spatial autocorrelated Getis-Ord *Gi z*-score values. A color version of this figure is available online

4.1 Pattern A, north/south

The map for the variable *grandpa/grandfather* displays a clear distinction between the usage of the terms in Northern and Southern California. In particular, the less formal realization for the concept 'the father of one's father or mother' is relatively more frequent in Northern California, while the more formal one is relatively more frequent in the lower part of Southern California, with a few weak outliers in the central part of the state (see Fig. 4).

With the variable *folks/parents* (Fig. 8), the scheme resembles the one in *grandpa/grandfather*. In fact, the less formal realization for the concept 'a person's father and mother' *folks* is more common in the north, while the more formal realization is more common in the south.

4.2 Pattern B, urban/rural

The variables *dad/father* and *mom/mother* reveal a clear usage distinction in written online newspaper

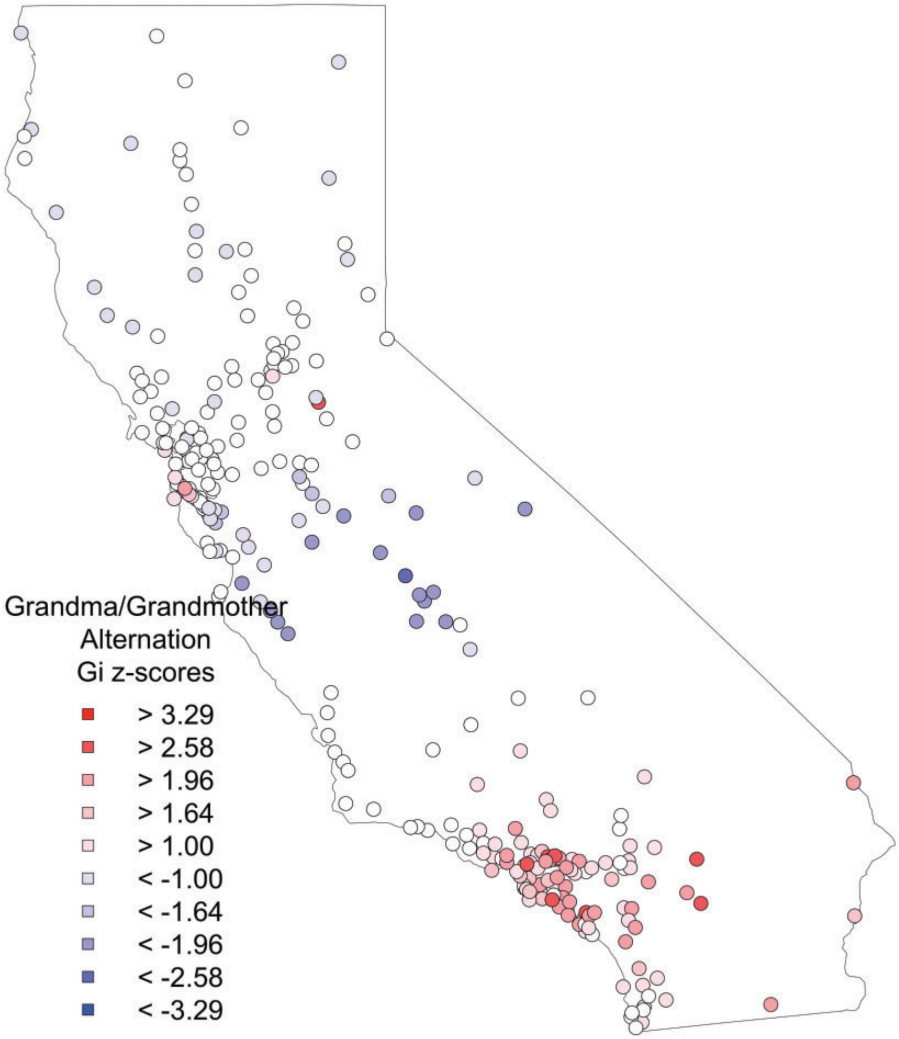

Fig. 5 Probability of *Grandma* relative to *Grandmother*: map of local spatial autocorrelated Getis-Ord *Gi z*-score values. A color version of this figure is available online

language between urban and rural California environments. In particular, for the variable *dad/father*, the variant *dad* is relatively more common in the metropolitan areas around San Francisco and Los Angeles; the variant *father* is relatively more common in the central and northern rural parts of California (Fig. 2).

For the variable *mom/mother*, the term *mom* is relatively more frequent in the San Francisco Bay region and in the Los Angeles area; the term *mother* is relatively more frequent in the more rural eastern part of Northern California (Fig. 3).

For the variable *grandma/grandmother*, the term *grandma*, which is the less formal realization for the concept 'the mother of one's father or mother' is relatively more frequent in Southern California, especially in the Los Angeles urban area, as well as in San Francisco and in the Silicon Valley; the more formal term *grandmother* is relatively more frequent in Northern California, with a prevalence in the

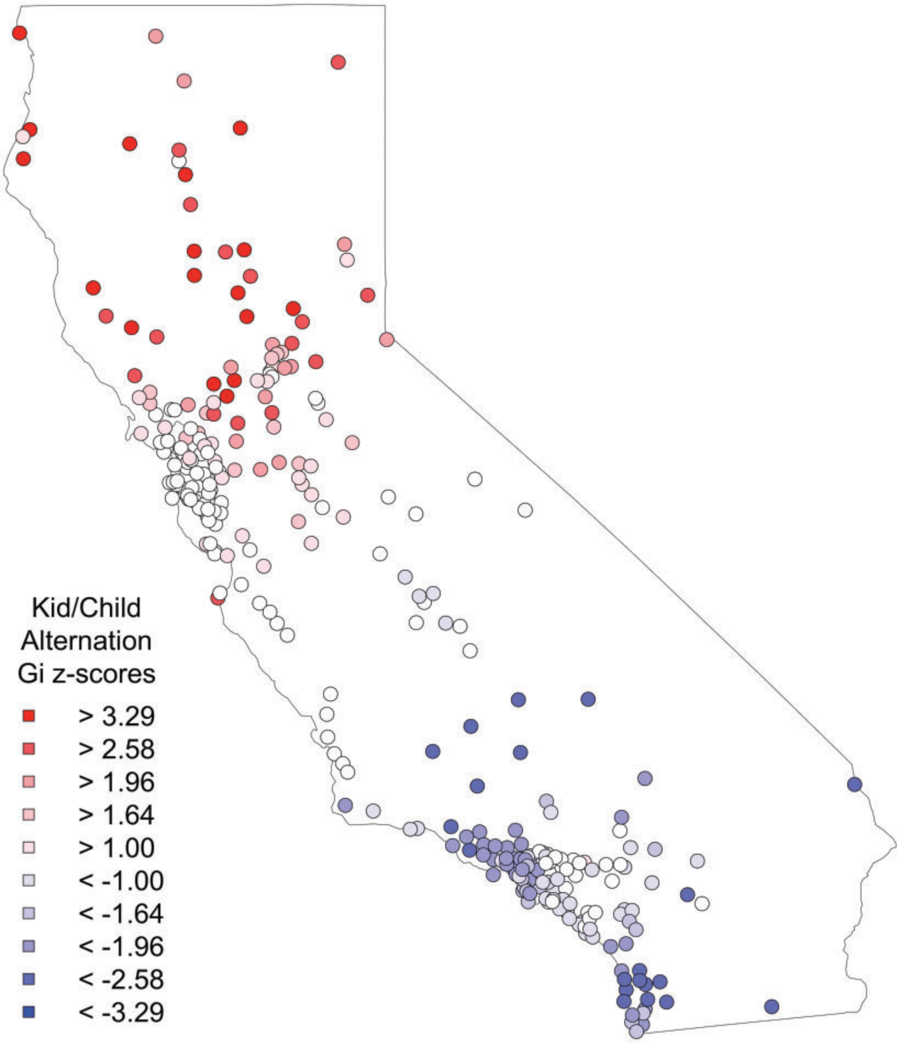

Fig. 6 Probability of *Kid* relative to *Child*: map of local spatial autocorrelated Getis-Ord *Gi* *z*-score values. A color version of this figure is available online

central part of the state, as well as on the coast of Northern California (see Fig. 5).

Finally, although *folks/parents* displays dialect patterns mainly on a north/south dimension, the variable has a notable strong correlation in the urban area of Greater Los Angeles (Fig. 8).

4.3 Pattern testing

In order to verify the two identified patterns and test them with further data, we compared the six distributional patterns for the lexical variables under examination in this article with the distributional patterns for another set of eight variables, namely contraction rate variables (*hasn't/has not*, *haven't/have not*, *doesn't/does not*, *don't/do not*, *wasn't/was not*, *weren't/were not*, *couldn't/could not*, *won't/will not*). For similarity reasons, data for this comparison were retrieved and analyzed following the same criteria as the ones we detailed in Sections 2 and 3.

Notably, the variables *doesn't/does not*, *don't/do not*, *couldn't/could not*, and *won't/will not* followed a north/south structure as in Pattern A (Section 4.1;

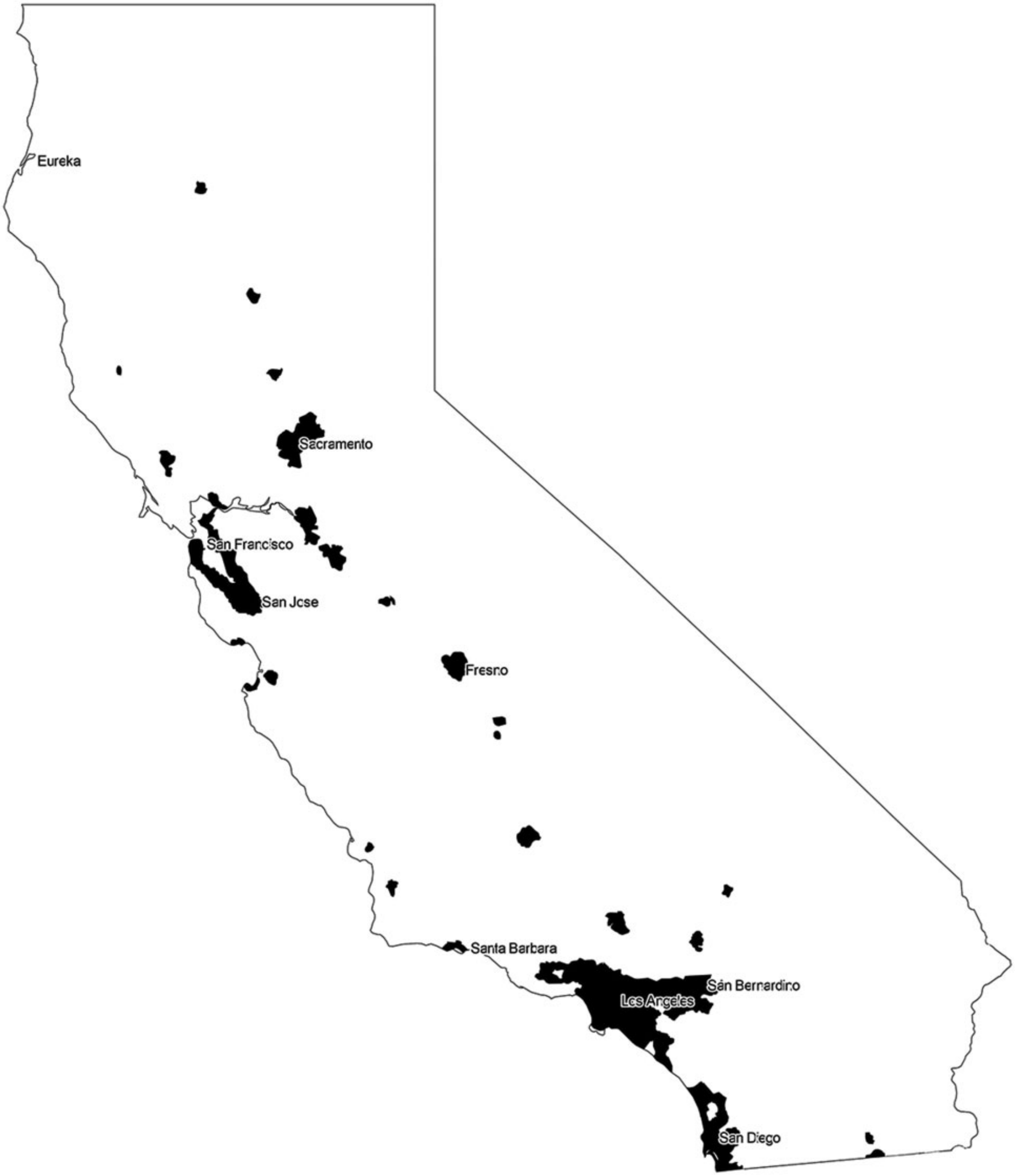

Fig. 7 Population distribution in California (Source: 2010 US census)

Fig. 9), and the variables *hasn't/has not*, *haven't/have not*, *wasn't/was not*, and *weren't/were not* followed an urban/rural structure as in Pattern B (Section 4.2; Fig. 10).

5 Discussion

As an attempt to answer Labov's (1972) call[5] for a quantification of the dimension of style, situating

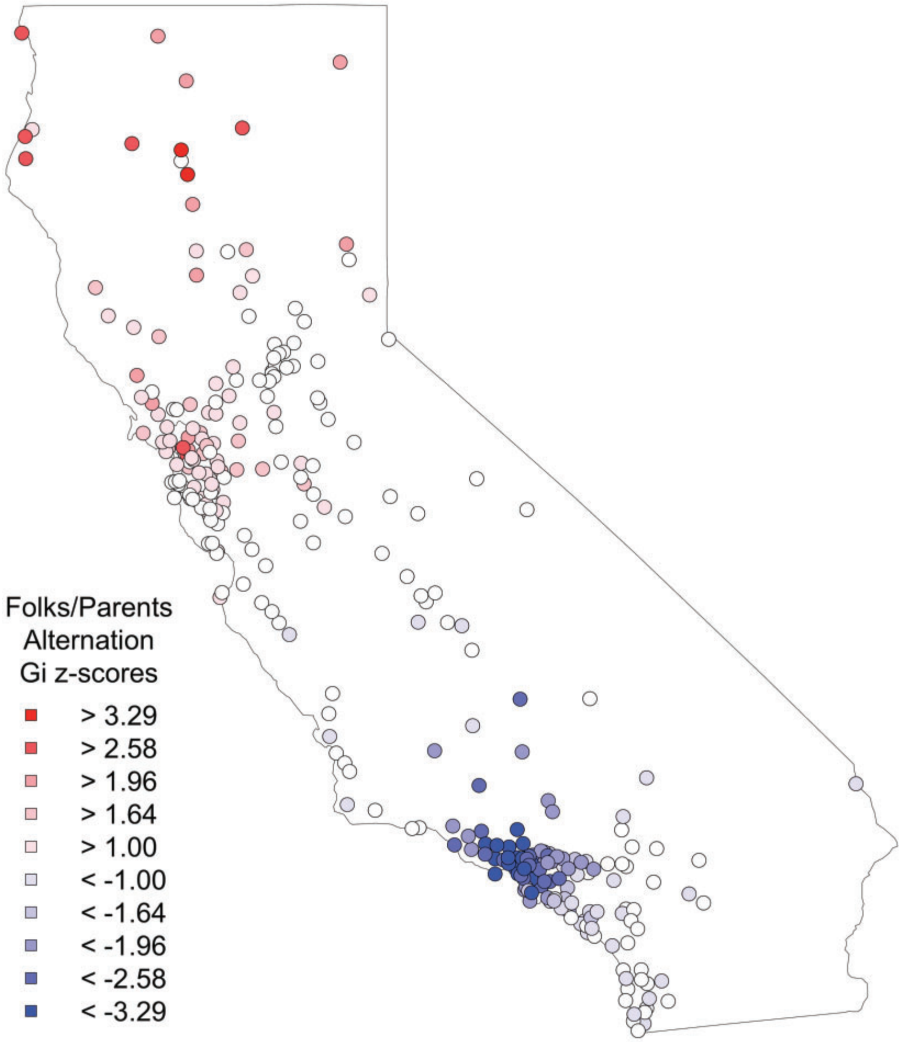

Fig. 8 Probability of *Folks* relative to *Parents*: map of local spatial autocorrelated Getis-Ord *Gi z*-score values. A color version of this figure is available online

the quantification on a geographical level, the analysis of distribution of the selected lexical variables formed by word alternations on different levels of formality brought to the conclusion that written formality in the English language is regionally patterned, as Grieve's (2011) analysis of American English demonstrated before. Therefore, regional linguistic variation on a formality level exists in written Standard California English. In particular, two very strong patterns of variation emerged for California English, namely an urban/rural dimension and a north/south dimension.

A historical motivation underlies the language usage distinction between the north and the south of California. In fact, in the mid-nineteenth century, while Northern California was growing rapidly as a consequence of the Gold Rush, Southern California continued to be a pastoral Hispanic region until the 1880s, when, with the development of irrigation and the aqueduct

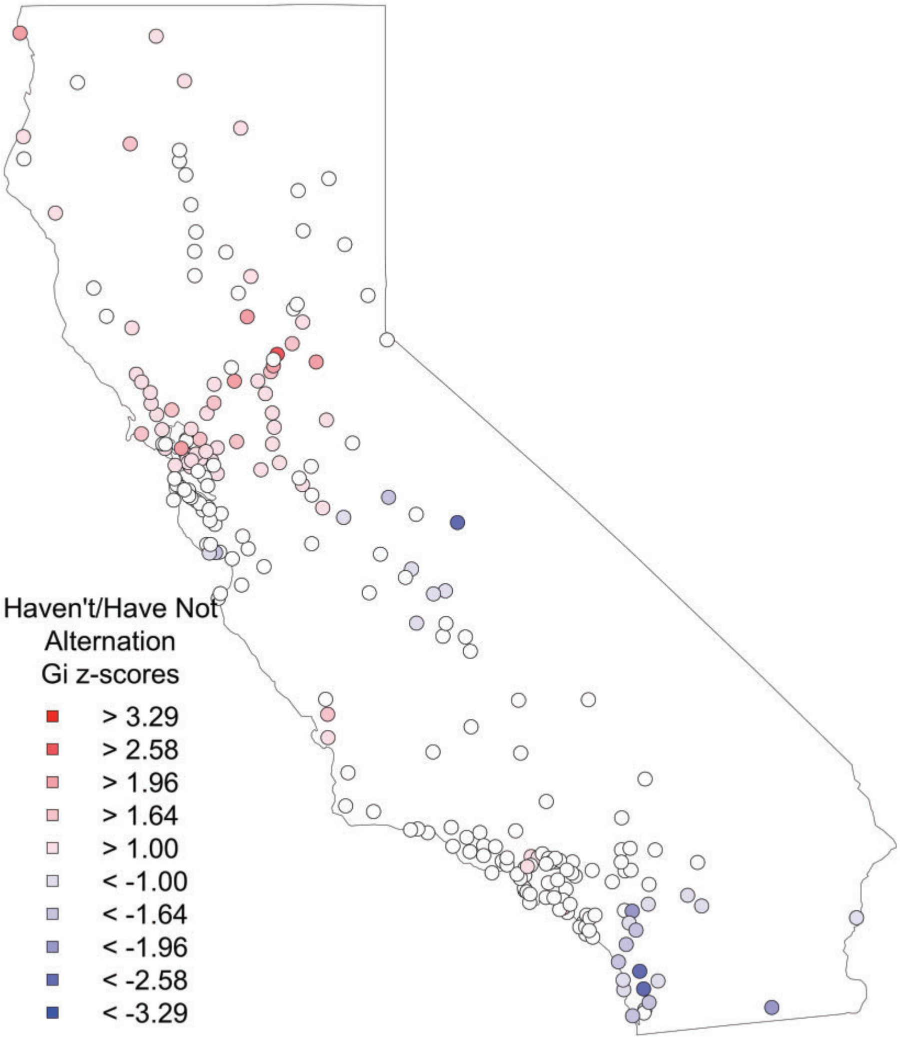

Fig. 9 Pattern A (north/south) testing: Probability of *Haven't* relative to *Have Not*: map of local spatial autocorrelated Getis-Ord *Gi z*-score values. A color version of this figure is available online

system, the Imperial Valley saw the increase of the farm population. The year-round favorable weather and the relaxed lifestyle drew people to Southern California, incrementing the real estate industry (Starr and Procter, 1973; Hayes, 2007). While San Francisco was the most populated city in the state until 1880, Los Angeles grew considerably in the following years and became three times as big as San Francisco in 1950.[6] Therefore, the residents of Northern California have been settled longer than the residents of Southern California. The different use of the language between north and south that is evidenced in this research can be a result of the historical settlement patterns. This study confirms what Reed pointed out 60 years ago: through the example of the distribution of the term *chesterfield*, Reed provided evidence that a north/south dialect distinction already existed in California in the 50s (Reed and Bradley, 1954).

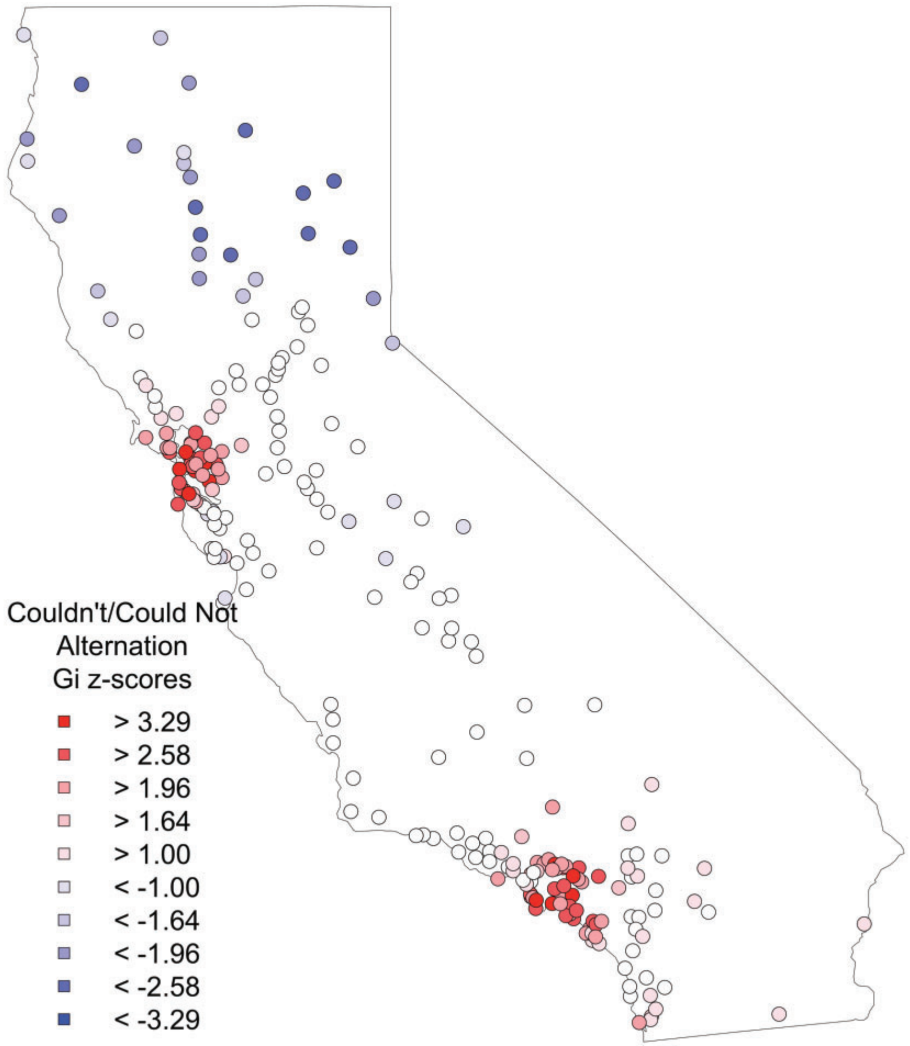

Fig. 10 Pattern B (urban/rural) testing: probability of *Couldn't* relative to *Could Not*: map of local spatial autocorrelated Getis-Ord *Gi z*-score values. A color version of this figure is available online

The straightforward language distinction between the rural and the urban areas of California is based on the socioeconomical structure, where a metropolis influences people's lifestyles in a different way if compared to what has an impact on people's habits in the agricultural cities. Language in general, included written language as demonstrated here, is gradually adapting toward informality, a 'shift to a more speech-like style' that Geoffrey Leech defines as 'colloquialization' (Leech et al., 2009: 239). The urban sprawls are normally motors for innovation, and they are also in the linguistic context, while the slower-paced countryside is reluctant to adopt changes. In fact, the results of this research show that metropolitan areas tend toward more informal language than rural areas. Notably, although California encompasses five urban agglomerations (Greater Los Angeles, the San Francisco Bay, San Diego-Tijuana, Greater Sacramento, and Metropolitan Fresno), according

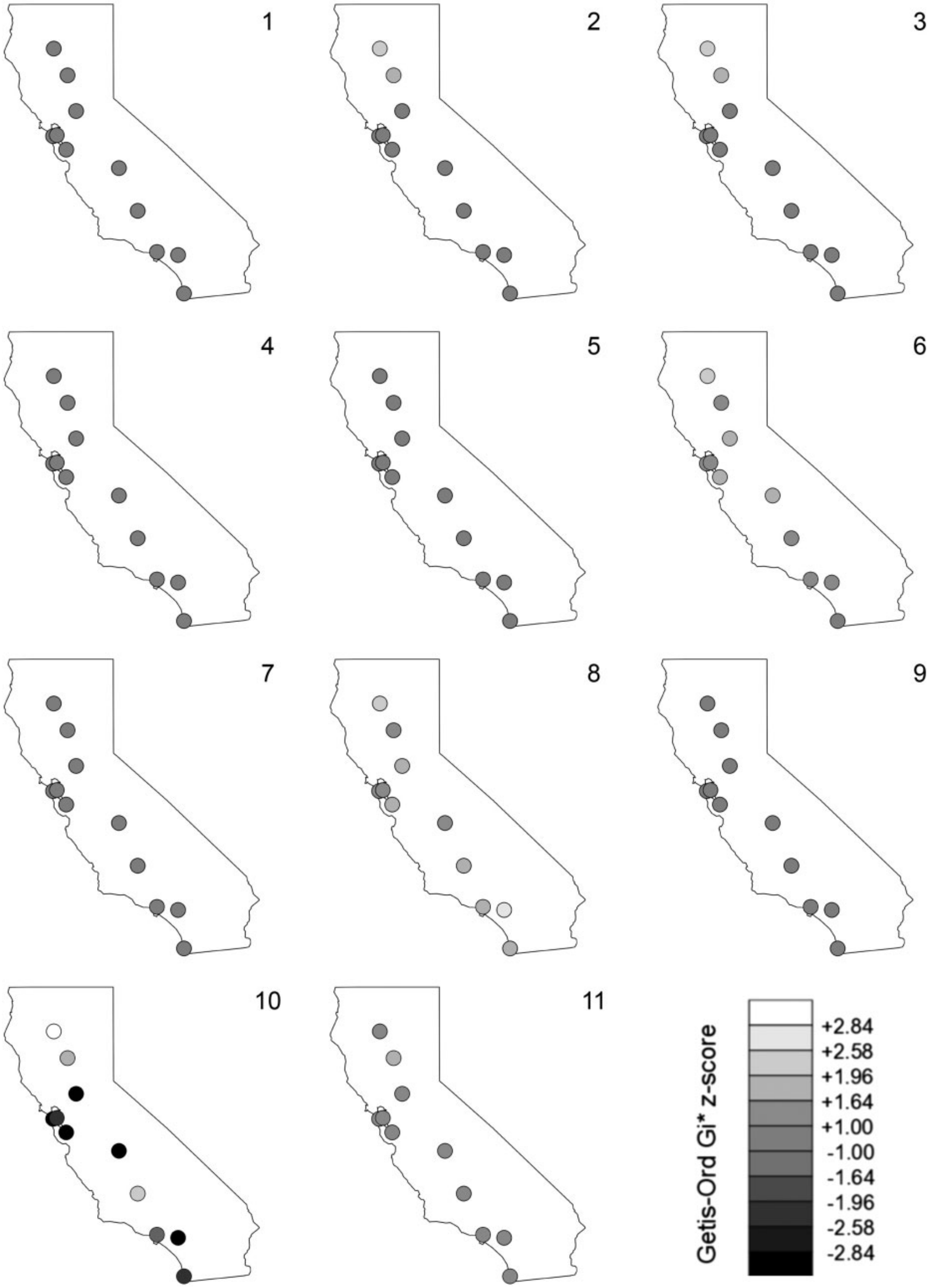

Fig. 11 Reproduction of Grieve's contraction measures maps—California Section (Source: Grieve, 2011): 1. Be Not Contraction; 2. Do Not Contraction; 3. Have Not Contraction; 4. Modal Not Contraction; 5. Be Contraction; 6. Have Contraction; 7. Modal Contraction; 8. Them Contraction; 9. To Contraction; 10. Non-Standard 'Not' Contraction; 11. Double Contraction

to this research, only two of those agglomerations are involved in a different use of the language compared to the rest of the state. In particular, the two urban areas that emerge from this study are Greater Los Angeles and the San Francisco Bay. Greater Los Angeles and the San Francisco Bay are in fact different from the rest of the California metropolitan areas from a variety of points of view. To pick one, the gross domestic product (GDP) of Greater Los Angeles and the San Francisco Bay is much higher than the GDP of any other California urban agglomerations.

A comparison between the results of the present study and previous research show that Grieve's (2011) arguments align to this study. The two studies are similar on the basis of goals, methods, and coverage of California. In fact, the goal of both this study and Grieve's study is to analyze language variation in written Standard American English as appears in online newspapers, and the methods used are somewhat comparable, if not in the data collection, at least in the statistical analyses. Both studies analyze English dialect variation on a quantitative basis, and both provide dialect maps. Also, although encompassing the US, Grieve's coverage includes California, which is also the state surveyed here. However, a comparison with Grieve's results is possible to a limited extent. The comparability is limited due to the different number of observations, the different geographical zoom of the two studies, and the different nature of the analyzed variables. With regards to the number of observations, Grieve's sample for California contains 11 cities, while this study analyzes 273 locations. As for the geographical zoom of the two studies, Grieve's survey focuses on the 48 contiguous USA, while this research focuses on only one out of those 48 states. Also, the variables analyzed by Grieve are full grammatical forms versus contracted grammatical forms, such as the alternation in writers' choice between *is not* and *isn't*, based on the assumption that contractions are prevalent in informal writing, while they tend to be avoided in more formal writing; this study analyzes two lexical forms for each onomasiological concept, based on the assumption that one lexical form is used in a more formal way whereas the other lexical form is used in a less formal way. Finally, the newspaper sample is limited to letters to the editor in Grieve's survey, while the sample for this research encompasses the whole archive that online newspapers make available.

Six out of the eleven contraction variables analyzed by Grieve witness variation in California (features 2, 3, 6, 8, 10, 11, Fig. 11), four of which display very weak patterns (features 2, 3, 6, 11, Fig. 11). The most successful variable in terms of regional variation in California is the *non-standard 'not' contraction* (feature 10, Fig. 11). In particular, the *non-standard 'not' contraction* variable presents great variation from the upper part of Northern California (Redding and Chico) to the San Francisco Bay Area, Sacramento, and Fresno; moreover, the *non-standard 'not' contraction* variable behaves similarly in the San Francisco Bay Area, Sacramento, Fresno, Los Angeles, Riverside, and San Diego, while in Bakersfield the variable displays a different behavior, more similar to the one detected in the northern rural area. It should be noted that these two different pattern regions relative to the results obtained from this variable are both rural; one cluster of observations is in the north, while one other observation is in the south. The patterns resulting from Grieve's survey for the *non-standard 'not' contraction* variable seem comparable to the results of the study reported in this article.

The territory under investigation could be expanded in a future study. For example, once the entire US territory is surveyed for patterns of lexical formality in written Standard American English, it would be interesting to compare lexical results to contraction rate patterns and to previously established general dialect patterns in formality variation in American English.

Acknowledgments

Many thanks to Jack Grieve and Thomas Wielfaert for their comments on the article.

References

Asnaghi, C. (2013). *An Analysis of Regional Lexical Variation in California English Using Site-Restricted*

Web Searches. Joint Ph.D. Dissertation, Università Cattolica del Sacro Cuore and University of Leuven, Milan, Italy and Leuven, Belgium.

Atwood, E.B. (1962). *The Regional Vocabulary of Texas*. Austin, TX: University of Texas Press.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge, UK: Cambridge University Press.

Bright, E. S. (1971). *A Word Geography of California and Nevada*, Volume 69. Berkeley, California: University of California Press.

Brooke, J., Wang, T., and Hirst, G. (2010). Inducing Lexicons of Formality from Corpora. In Bel, N., Daille, B., and Vasiljevs, A. (eds), *Proceedings of the LREC 2010 Workshop*. Valletta, Malta, pp. 17–22.

Bucholtz, M., Bermudez, N., Fung, V., Edwards, L., and Vargas, R. (2007). Hella nor cal or totally so cal? The perceptual dialectology of California. *Journal of English Linguistics*, **35**: 325–52.

Carver, C. M. (1987). *American Regional Dialects: A Word Geography*. Ann Arbor, MI: University of Michigan Press.

Cassidy, F. G. and Hall, J. H. (eds), (1985). *Dictionary of American Regional English*. vol. 1A–C. Cambridge, MA: Belknap Press of Harvard University Press.

Cassidy, F. G. and Hall, J. H. (eds), (1991). *Dictionary of American Regional English*. vol. 2D–H. Cambridge, MA: Belknap Press of Harvard University Press.

DeCamp, D. (1971). The pronunciation of English in San Francisco. In Williamson, J. V. and Burke, V. M. (eds), *A Various Language: Perspectives on American Dialects*. New York: Holt, Rinehart and Winston, Inc.

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. vol. 27. Hoboken, NJ: Wiley-Blackwell.

Ferguson, C. A. (1994). Dialect, register and genre: working assumptions about conventionalization. In Biber, D. and Finegan, E. (eds), *Sociolinguistic Perspectives on Register*. Oxford, UK: Oxford University Press, pp. 15–30.

Geeraerts, D., Grondelaers, S., and Speelman, D. (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat. Een Onderzoek Naar Kleding- En Voetbaltermen*. Amsterdam: Meertens Instituut.

Grieve, J. (2009). *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English* Ph.D. thesis, Flagstaff, Arizona: Northern Arizona University.

Grieve, J. (2011). A regional analysis of contraction rate in written standard American English. *International Journal of Corpus Linguistics*, **16**(4): 514–46.

Grieve, J., Asnaghi, C., and Ruette, T. (2013). Site-restricted web searches for data collection in regional dialectology. *American Speech*, **88**: 413–40.

Grieve, J., Speelman, D., and Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, **23**: 193–221.

Hagiwara, R. (1995). *Acoustic Realizations of American/r/as Produced by Women and Men*, vol. 90. Los Angeles: Phonetics Laboratory, Department of Linguistics, UCLA.

Hall, J. H. (ed.) (2002). *Dictionary of American Regional English*, vol. 4. P-Sk. Cambridge, MA: Belknap Press of Harvard University Press.

Hall, J. H. (ed.) (2012). *Dictionary of American Regional English*, vol. 5. Sl-Z. Cambridge, MA: Belknap Press of Harvard University Press.

Hall, J. H. and Cassidy, F. G. (eds), (1996). *Dictionary of American Regional English*, vol. 3I-O. Cambridge, MA: Belknap Press of Harvard University Press.

Hall, J. H. and von Schneidemesser, L. (eds), (2013). *Dictionary of American Regional English, Volume 6: Contrastive Maps, Index of Entry Labels, Questionnaire, and Fieldwork Data*. Cambridge, MA: Belknap Press of Harvard University Press.

Hall-Lew, L. (ed.) (2010). *Ethnicity and Phonetic Variation in a San Francisco Neighborhood* Ph.D. thesis, Stanford University.

Hayes, D. (2007). *Historical Atlas of California: With Original Maps*. Berkeley/Los Angeles/London: University of California Press.

Hinton, L., Moonwomon, B., Bremner, S., Luthin, H., Clay, M. V., Lerner, J., and Corcoran, H. (1987). It's not just the valley girls: a study of California English. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, **vol. 13.**

Hubert, L. J., Golledge, R. G., and Constanzo, C. M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, **13**: 224–33.

Kennedy, R. and Grama, J. (2012). Chain shifting and centralization in California vowels: an acoustic analysis. *American Speech*, **87**(1): 39–56.

Kretzschmar, W. (2009). *The Linguistics of Speech*. Cambridge, UK: Cambridge University Press.

Kurath, H. (1949). *A Word Geography of the Eastern United States*. Ann Arbor, MI: University of Michigan Press.

Labov, W. (1972a). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1972b). Some principles of linguistic methodology. *Language in Society*, **1**: 97–120.

Labov, W., Ash, S., and Boberg, C. (2006). *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

Leech, G., Hundt, M., Mair, C., and Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge, UK/New York: Cambridge University Press.

Moonwomon, B. (1991). *Sound Change in San Francisco English*. Ph.D. thesis, Berkeley: University of California.

Moran, P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society Series B (Methodological)*, **10**(2): 243–51.

Odland, J. (1988). *Spatial Autocorrelation*. London: Sage Publications.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, **27**(4): 286–306.

Paradis, E. (2009). Moran's Autocorrelation Coefficient in Comparative Methods http://cran.r-project.org/web/packages/ape/vignettes/MoranI.pdf. (accessed 1 September 2013).

Podesva, R. J. (2011). The California vowel shift and gay identity. *American Speech*, **86**(1): 32–51.

Reed, D. W. and Bradley, F. W. (1954). *Eastern Dialect Words in California* (Special Issue Of: American Speech: A Quarterly of Linguistic Usage ed.), vol. 21. Ann Arbor, MI: Publication of the American Dialect Society.

Reed, D. W. and Metcalf, A. A. (1952). *Linguistic Atlas of the Pacific Coast*. Microfilm: Bancroft Library of the University of California at Berkeley.

Ruette, T., Speelman, D., and Geeraerts, D. (2011). Measuring the lexical distance between registers in national varieties of Dutch. In Soares da Silva, A. T., Torres, A., and Gonçalves, M. (eds), *Línguas Pluricêntricas. Variação Linguística e Dimensões Sociocognitivas*. Braga, Portugal: Publicações da Faculdade de Filosofia, Universidade Católica Portuguesa, pp. 541–54.

Sawada, M. (2004). *Global spatial autocorrelation indices – Moran's I, Geary's C and the General Cross-Product Statistic* Research paper from the Laboratory for Paleoclimatology and Climatology at the University of Ottawa. http://www.lpc.uottawa.ca/publications/moransi/moran.htm.

Speelman, D., Grondelaers, S., and Geeraerts, D. (2003). Profile-based linguistic uniformity as a generic method for comparing language varieties. *Computers and the Humanities*, **37**: 317–37.

Starr, K. and Procter, B. (1973). Americans and the California dream, 1850–1915. *History: Reviews of New Books*, **1**(9): 201.

Szmrecsanyi, B. (2014). Forests, trees, corpora, and dialect grammars. In Szmrecsanyi, B. and Wälchli, B. (eds), *Aggregating Dialectology, Typology, and Register Analysis*. Berlin: De Gruyter, pp. 89–112.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46**(2): 234–40.

Trudgill, P. (1999). Standard English: what it isn't. In Bex, T. and Watts, R. J. (eds), *Standard English: The Widening Debate*. London: Routledge, pp. 117–28.

Waksler, R. (2000). *A Hella New Specifier* ling.ucsc.edu/Jorge.

Notes

[1] The complete list of 334 online newspapers, including the location where the selected newspaper are based, the communities and topics they cover, the frequency of their publication, the circulation, and other notes can be found in Asnaghi, 2013.

[2] Only in this case the plural form was searched (*folks*/*parents*) to avoid ambiguous cases that would have occurred in the case of a search for *folk*/*parent*, such as hits for *folk* meaning 'folk music' and 'folk art'.

[3] A brief survey was sent to a sample of California newspaper editors. About 30 editors replied, confirming that: 'Almost all [journalists] are local residents' (Becky O'Malley, Editor, *Berkeley Daily Planet*) or 'We [i.e. the journalists] are all residents of the city' (Judi Bowers, Editor, *Big Bear Grizzly*).

[4] San Francisco Chronicle, 14 December 2001, http://www.sfgate.com/opinion/letterstoeditor/article/LETTERS-TO-THE-EDITOR-2840507.php, retrieved on 17 July 2014.

[5] ''The most immediate problem to be solved in the attack on sociolinguistic structure is the quantification of the dimension of style'' (Labov, 1972: 245).

[6] Information retrieved from quickfacts.census.gov on 7 December 2012.