User Review Sites as a Resource for Large-Scale Sociolinguistic Studies

Dirk Hovy  
Center for Language Technology  
University of Copenhagen, Denmark  
Njalsgade 140  
dirk@cst.dk

Anders Johannsen  
Center for Language Technology  
University of Copenhagen, Denmark  
Njalsgade 140  
ajohannsen@hum.ku.dk

Anders Søgaard  
Center for Language Technology  
University of Copenhagen, Denmark  
Njalsgade 140  
søgaard@hum.ku.dk

ABSTRACT

Sociolinguistic studies investigate the relation between language and extra-linguistic variables. This requires both representative text data and the associated socio-economic meta-data of the subjects. Traditionally, sociolinguistic studies use small samples of hand-curated data and meta-data. This can lead to exaggerated or false conclusions. Using social media data offers a large-scale source of language data, but usually lacks reliable socio-economic meta-data. Our research aims to remedy both problems by exploring a large new data source, international review websites with user profiles. They provide more text data than manually collected studies, and more meta-data than most available social media text. We describe the data and present various pilot studies, illustrating the usefulness of this resource for sociolinguistic studies. Our approach can help generate new research hypotheses based on data-driven findings across several countries and languages.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms
Human Factors; Experimentation

Keywords
Language-analysis techniques; Multi-lingual and cross-lingual analysis and mining; Social science research based on social media; Insights from natural-language analysis of social media; Novel applications

1 Introduction

Sociolinguistic studies investigate the relation between language use and extra-linguistic socio-economic variables, such as regional origin, gender, social class, or age [17, 29, 25, 27, 9]. These studies thus require both a representative corpus of text data and meta-data about the socio-economic variables. Traditionally, these studies have relied on individual interviews and questionnaires and their manual transcriptions to collect data and meta-data. Due to the effort in curating them, the resulting corpora are often of moderate size, and sometimes include less than five subjects [27]. However, small sample sizes may lead to false research findings, as elaborated on in [8] (albeit for neuro-sciences). Sociolinguistic studies, in other words, often lack statistical power to either establish relationships between language use and socio-economic variables (because they are limited to detect only large effects), or are liable to present exaggerated claims (if they report spurious effects due to a small sample size).

Web data, on the other hand, presents a source of large amounts of text data. In addition, a lot of web data is personalized text (i.e., not canonical), such as blogs and other social media texts. I.e., they represent actual language use, rather than a prescriptively normed (standard) variant. This allows sociolinguistic studies with sufficient statistical power, but requires a web-mining approach. Indeed, the natural language processing (NLP) community has recently shown a growing interest in sociolinguistic research questions, applying modern data-driven methods to social media data. This interest was in part driven by the availability of some meta-data (e.g., location and time) in social media platforms such as Twitter [14, 13, 12, 5].

The meta-data for social media, however, is partial at best and often unreliable, and so most previous work has only considered regional origin and time as extra-linguistic variables. This severely limits the types of sociolinguistic research questions that can be addressed, because most traditional studies are concerned with socio-economic factors such as age, gender, or class, which are not present in social media meta-data. To remedy this problem of incomplete meta-data, a whole branch of previous work has thus focused on building predictive models for age and gender [7, 4, 10] to add user information. Some social media data sets, like Facebook, contain more meta-data, but are difficult or impossible to obtain.

This paper proposes a novel source of data for sociolinguistic studies, namely user-review sites. We show that these sites contain textual data (reviews) that is linked to various extra-linguistic information, namely age, gender, and location of the reviewer, as well as time-stamps. The language in the reviews, while not as informal as e.g., Twitter, is much
less canonical than newswire, and thus likely to reflect the socio-economic background of the user. More importantly, some user-review sites contain up to millions of user reviews and thus provide the statistical power to study the influence of socio-economic variables on language use.

In our case, we extract data from the website of Trustpilot\(^1\). On Trustpilot, users review online and brick-and-mortar companies they shopped at, and leave a one to five star rating, as well as a written review. The website offers reliable information about the time the review was posted, so it can be used to track some diachronic development. In addition, users often supply information about their location, age, and gender. The data is available for 24 countries, using 13 different languages (Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish), and thus contains comparable text across socio-economic variables in several languages, spanning several years.

We explore this new data source and its usefulness to test sociolinguistic hypotheses with sufficient statistical power.

**Our contributions**

We show how publicly available user-review sites can be used for large-scale sociolinguistic experiments. The size of the corpus allows for statistically more powerful analyses than manually-collected corpora. The sites also contain more meta-data (information about the speaker, including age, gender, and location) than social media data.

We describe the process of collecting data from the Trustpilot website, and how to augment missing information. The scripts for harvesting and preprocessing the corpora are made publicly available.\(^2\)

We analyze the resulting corpus with respect to representativeness for all age ranges, both genders, and various regions. Since the data consists of user reviews, we also study the distribution of ratings across socio-economic variables.

We then present a series of pilot studies across several languages. The studies are based on hypotheses found in sociolinguistic literature and include gender-specific lexical differences, the distribution of regional/dialect markers, the decline of grammatical constructions, and spelling variations. Our results indicate that user-review sites can provide informative data for large-scale sociolinguistic studies.

## 2 Related Work

Sociolinguistic and variational studies have a long tradition, correlating speakers’ linguistic traits with socio-economic background and variables such as origin, social class, age, and gender.

Labov’s seminal studies \([19]\) investigated the influence of regional origin and class on phonology. In recent years, sociolinguistic studies have investigated the influence of age \([27]\), gender \([17, 27]\), class \([24, 22, 23]\), and ethnicity \([9, 27]\). All of them find that language use is highly influenced by these factors, although a statistical correlation is sometimes tenuous or absent.

With the widespread availability of Twitter, recent studies in NLP have focused on exploring linguistic features such as phonological, lexical, and morphological variation, and how they are influenced by spatial resolution \([14, 13, 28, 12, 5]\), gender \([6]\), and ethnicity \([32]\). Some of these research questions, such as which phonological features influence written language on social media, are interesting on their own, and NLP offers tools for answering these questions \([13]\). Other sociolinguistic research questions coincide with important topics in NLP, e.g., language change, which challenges the robustness of the predictive models used in the field \([14, 30]\). While Twitter provides a large amount of constantly updated data, it is notoriously difficult to verify sociolinguistic traits of the speakers. Some inferences can be made by correlating linguistic differences to census data \([13]\).

In this work, we use a resource that provides both textual and meta-data.

There is also an active research area in predicting sociolinguistic traits of authors, mainly focusing on age and gender \([4, 10, 21]\) using information such as linguistic choice and names as input. \([26]\) presented a study where they caution that humans apparently use various other factors to assess a user’s age and gender. \([20]\) recently presented a relation-extraction inspired approach to inferring a user’s spouse, education, and job by using distant supervision from various social networks. While interesting, these “acquired” features are only secondary to language use.

Areal studies of language variation across time have existed systematically at least since the neo-grammarians. Georg Wenker collected 30k transcriptions of dialect variations in the 1880s. However, they do not include meta-information about the speakers themselves, other than the name of the transcription (local school teachers). The results were mapped out to show regional variations and have recently been published along with the questionnaires by \([29]\). Similar work, but with a focus on computer-aided clustering and mapping of Dutch dialects, has been undertaken by \([25]\). A more historical approach to data-driven areal linguistic studies and language phylogeny was presented by \([11]\).

Other studies that investigate linguistic variation, albeit across languages and stylistic variation, include \([16]\) and \([18]\).

## 3 Data collection

### 3.1 Data format

| **Users** | Name, ID, profile text, location (city and country), gender, year of birth |
| **Reviews** | Title, text, rating (1–5), User ID, Company ID, Date and time of review |
| **Company** | Name, categories (e.g., electronics), number of ratings, description, address, location (city and country) |

\(^1\)\url{http://trustpilot.com}  
\(^2\)\url{http://www.lowlands.dk/results}
In order to add gender information when it is not supplied, we collect statistics for each occurrence of the name where gender information is available. This provides us with a distribution over genders for each name. If a name is found with sufficient frequency in our data, and predominantly for one gender, we propagate this information to all occurrences of the name that lack gender information. In our experiments, we used a gender-ratio of 0.95 (name occurs with one gender at least 95% of the time) and a minimum frequency of 3 (name appears at least 3 times with gender information in the data). Since the gender of some names is country-specific (e.g., Simone is a male name in Italy, but female in Germany or France), we repeat this step for each country separately. This augmentation step typically doubles the amount of available data with gender information for any language, while minimizing the risk of introducing false positives.

To add geographical information, we use the Geonames database\(^3\) to attach latitude and longitude information to the user. This allows areal distribution analysis of the data. Since the place names are entered as free text by the user, we apply various heuristics to get an acceptable match percentage. We remove short suffixes (e.g., “k” from “københavn”), spaces, and punctuation, spell out abbreviations, and try to correct misspellings with an edit distance of 1 from a known place name. Latitude and longitude values allow us to precisely place the user review on a map.

One problem with locations is the ambiguity of some place names, say “Nykøbing”, “Neustadt”, or “Springfield”. In many cases there is a “canonical” town with this name in a country, typically the largest one (one such example is “Kastrup” in Denmark, which refers to two cities on Zealand, as well as several villages in Jutland, but is generally understood to refer to the one close to the eponymous airport in Copenhagen). We determine the canonical location using a set of heuristics based on a population database. When the heuristics fail (e.g. when towns are of similar size, or more than one town has a population above a certain threshold), the location is left out of the analysis. Another, for now unanswered, problem is that the stated location most likely refers to the current residence of the user, and thus provides no information about their birthplace or where they grew up (typically important variables in variational linguistics).

### 4 Representativeness

Since the user information is given voluntarily and not verified, some birth years are presumably spurious (it seems unlikely that 2-year-olds or 110-year-olds review businesses on the Internet). In our experiments involving age, we thus restrict ourselves to the range from 16 to 80.

Apart from the outliers, the age distributions for both genders follow a reasonable distribution (see Figure 1). Additionally, the median age in our data is typically close to the country’s median value according to the CIA World Factbook\(^3\) (see Table 3), deviating slightly to both sides. For women, our mean-absolute error (MAE) over all countries is 0.74, for men 0.44. This indicates that the distributions are reasonably representative for the generations as a whole.

For gender, however, the differences are larger. Within each country, the gender distributions look similar, but there

---

\(^3\)www.geonames.org
In all studies, our methodology is to use the reviews to extract lexical variation and compute correlations of the variants with a socio-economic variable, i.e., gender, age, or region. E.g., we count how often a certain word X and its variant Y are used by a certain socio-economic group, and how much the variation is correlated with the group. In the case of gender, we compute $\chi^2$ over the contingency matrix. For correlation of variants with age and region, we use the Spearman coefficient. We are only interested in highly significant correlations, so we set our $p < 0.001$.

6 General

6.1 Discovering gender-specific words

If we already have a pre-defined sociolinguistic hypothesis, it is easy to verify it with enough data. However, such hypotheses require already-existing knowledge about the target variables. This limits what we can explore, and can also introduce bias from the researcher, by restricting the alternatives that are considered. One of the advantages of a large data collection is the ability to discover variations in a purely data-driven manner. [15] presented a data-driven approach to discover ethnicity-specific words using $l_1$-regularization. In this pilot study, we look at gender, and use a simpler method ($tf-idf$) to learn salient expressions.

In order to find the most gender-specific words, we separate the reviews from each language into two sub-corpora, based on the reviewers stated or inferred gender. We then compute $tf-idf$ on each sub-corpus separately and normalize them. $tf-idf$ weights content words that occur in fewer documents higher than common words (prepositions, forms of to be, etc.). We can now compare the relative $tf-idf$ scores for both genders and compute the difference by subtracting one from the other. This gives us a gradient, with predominantly “female” words on one and predominantly “male” words on the other end of the spectrum. Table 4 gives an overview of the 10 most gender-specific words for each language. In general, women tend to express satisfaction (“tilfreds”, “satisfaite”, “content” “pleased”) in all languages (except German), while men seem to talk more about problems (“problemer”, “probleme”) and agreement (“aftalt”, “conforme”). For French, where adjectives in sentences like “I am ADJ” are inflected according to the gender of the speaker, we see that the word for “satisfied” appears in both top 10. For both varieties of English, we find that certain words are associated more with women than with men (“loved”, “pleased”).

We do not attempt to interpret these findings any further, but point out that this might be a valuable approach to collect information about gender-specific differences in a data-driven way.

6.2 Emoticons, age, and gender

Emoticons have traditionally been seen as markers of adolescent language. We thus expect to see emoticons of all sorts used more by younger speakers.

We correlate emoticon usage with age and gender. We define emoticons to be a combination of eyes (such as ; ; ; or X), nose (~, or none), and mouth (\, \, [, *, etc). Overall, we check for 66 variations and record the ratio of emoticons in the overall number of words used in the category.
Table 3: Mean and median age for both genders over all countries, in data and official median

<table>
<thead>
<tr>
<th></th>
<th>WOMEN mean</th>
<th>WOMEN median</th>
<th>WOMEN off. median</th>
<th>MEN mean</th>
<th>MEN median</th>
<th>MEN off. median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denmark</td>
<td>38.80</td>
<td>38</td>
<td>41.6</td>
<td>39.07</td>
<td>38</td>
<td>39.8</td>
</tr>
<tr>
<td>France</td>
<td>42.03</td>
<td>41</td>
<td>41.2</td>
<td>41.92</td>
<td>41</td>
<td>38.2</td>
</tr>
<tr>
<td>Germany</td>
<td>40.64</td>
<td>40</td>
<td>45.3</td>
<td>38.97</td>
<td>38</td>
<td>42.3</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>44.51</td>
<td>45</td>
<td>41.5</td>
<td>43.87</td>
<td>43</td>
<td>39.4</td>
</tr>
<tr>
<td>United States</td>
<td>40.79</td>
<td>40</td>
<td>38.1</td>
<td>36.70</td>
<td>33</td>
<td>35.5</td>
</tr>
</tbody>
</table>

Figure 2: Regional coverage of our data. The figure compares geographical distribution of the population by region with the distribution in our dataset. Source: [2].

We find that for all countries, use of emoticons is strongly anti-correlated with age (Spearman $\rho = -0.99$ over all countries, see Table 5) at $p < 0.001$. In other words, as expected, the older a user is, the less likely they are to use emoticons.

For gender, we find that women use emoticons almost twice as often as men do (0.10% vs 0.18% of the respective words across all languages). The difference in the ratios is significant for all languages but French and German, although for the aggregated total of all languages, it is significant.

The presence or absence of a nose is a distinguishing feature of emoticon use, and seems to vary a lot. We suspect that this is correlated with user age. We select the most frequent emoticons (:-), ;-), and :-D) and collect counts over the occurrences with and without a nose.

We find that for all ages, the use of a nose is highly anti-correlated with age (at $p < 0.001$). Figure 3 shows the distribution over all languages, which shows a different interesting fact: there are two distributions at work. Use of a nose increases steadily with age, until at around 34, it overtakes the use of no nose. Given that there are fewer older
Table 4: The 10 most gender-specific words per language, based on relative tf-idf

<table>
<thead>
<tr>
<th></th>
<th>Denmark</th>
<th>France</th>
<th>Germany</th>
<th>UK</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>nemt</td>
<td>satisfaite</td>
<td>gerne</td>
<td>flowers</td>
<td>customer_service</td>
</tr>
<tr>
<td></td>
<td>bestille</td>
<td>contente</td>
<td>qualität</td>
<td>really</td>
<td>love</td>
</tr>
<tr>
<td></td>
<td>varer</td>
<td>suis</td>
<td>tolle</td>
<td>pleased</td>
<td>we</td>
</tr>
<tr>
<td></td>
<td>tilsfreds</td>
<td>produits</td>
<td>wir</td>
<td>thank_you</td>
<td>pleased</td>
</tr>
<tr>
<td></td>
<td>rigtig</td>
<td>chaussures</td>
<td>ware</td>
<td>definitely</td>
<td>really</td>
</tr>
<tr>
<td></td>
<td>handle</td>
<td>étiquettes</td>
<td>bestellen</td>
<td>lovely</td>
<td>petflow</td>
</tr>
<tr>
<td></td>
<td>dejligt</td>
<td>sont</td>
<td>sind</td>
<td>love</td>
<td>free_shipping</td>
</tr>
<tr>
<td></td>
<td>nem</td>
<td>ma</td>
<td>karten</td>
<td>quickly</td>
<td>thank_you</td>
</tr>
<tr>
<td></td>
<td>hjemmeside</td>
<td>été</td>
<td>bestellung</td>
<td>received</td>
<td>our</td>
</tr>
<tr>
<td></td>
<td>mine</td>
<td>spartoo</td>
<td>vielen_dank</td>
<td>improved</td>
<td>food</td>
</tr>
<tr>
<td>male</td>
<td>lovett</td>
<td>problème</td>
<td>keys</td>
<td>wife</td>
<td>these_guys</td>
</tr>
<tr>
<td></td>
<td>deres</td>
<td>rien</td>
<td>günstig</td>
<td>parts</td>
<td>parts_geek</td>
</tr>
<tr>
<td></td>
<td>problemer</td>
<td>bonne</td>
<td>service</td>
<td>deal</td>
<td>fast</td>
</tr>
<tr>
<td></td>
<td>alt</td>
<td>matériel</td>
<td>mmoga</td>
<td>product</td>
<td>partsgeek</td>
</tr>
<tr>
<td></td>
<td>aftalt</td>
<td>conforme</td>
<td>support</td>
<td>00</td>
<td>prices</td>
</tr>
<tr>
<td></td>
<td>virker</td>
<td>ldlc</td>
<td>weiterso</td>
<td>an</td>
<td>best</td>
</tr>
<tr>
<td></td>
<td>tiden</td>
<td>service</td>
<td>top</td>
<td>first_class</td>
<td>shipping</td>
</tr>
<tr>
<td></td>
<td>ok</td>
<td>sans</td>
<td>key</td>
<td>tyres</td>
<td>customer</td>
</tr>
<tr>
<td></td>
<td>pris</td>
<td>bon</td>
<td>guter</td>
<td>could_find</td>
<td>part</td>
</tr>
<tr>
<td></td>
<td>00</td>
<td>satisfait</td>
<td>gerne_wieder</td>
<td>best</td>
<td>parts</td>
</tr>
</tbody>
</table>

Table 5: Correlation of emoticons with age, and % of words for gender in different countries. Significance tests at $p < 0.001$

users and that the use of emoticons declines with age, this effect is likely to be even more distinct.

With respect to gender, we find that women tend to use the noseless variant significantly more than men, except for France, where the difference between genders is not statistically significant at the chosen level.

Table 6: Correlation of nose-use in emoticons for age and gender in different countries. All correlations significant at $p < 0.001$

6.3 Ratings, categories, gender, and age

Gender Apart from the linguistic data, we also have ratings associated with the reviews. Users rate the business on a 5-point Likert scale. We restrict ourselves to 1, 3 and 5-star ratings (borrowed from the sentiment analysis categories positive, negative, neutral) and record the distribution for each gender to see whether any differences emerge.

Table 7 shows the distributions of labels over gender and age. We find that men tend to vote slightly more negative than women, using fewer 5-star and more 1-star reviews. While the difference between genders on those two ratings
is statistically significant, the overall difference between the two distributions is not.

Similarly, people in the younger group are more likely to use negative ratings than people in the older group (see Table 7). While the differences are small, they show that there are measurable differences worth modeling.

The categories used to classify businesses show a very different behavior. Figure 4 shows that the categories are highly gender-specific. With the exception of Hotels and Fashion Accessories, the two distributions are almost opposites.

**Age** We also hypothesized that users of different age would exhibit different rating behavior. Again, we collected all available data. Rather than mapping a distribution for each age, we compute the average rating score.

While all ages have an average rating above 4, we find that older users tend to be more generous (or satisfied), which is reflected in a Spearman $\rho$ of 0.92, i.e., a strong correlation between user age and average rating. The correlation is significant at $p < 0.001$.

### Denmark

#### 7.1 Reflexive possessive pronouns

One of the distinguishing features of the western Danish dialect of Jutland is the missing distinction between the reflexive possessive pronouns *sin/sit* (“his/her own”) and the non-reflexives *hans/hendes/dens/dets* (“his”). I.e., there is no distinction between “He met his (own) wife” and “He met his (=someone else’s) wife”.

![Figure 5: Regional distribution of “sin/sit” (ratio of all pronouns per NUTS-2 region)](image)

In order to investigate this phenomenon in a data-driven way, we record the frequency of *sin/sit* and the joint frequency of all possessive pronouns. We then compute the ratio of the former in all pronouns. That is

$$r = \frac{\text{count}(\text{sin, sit})}{\text{count}(\text{sin, sit, hans, hendes, dens, dets})}$$

(1)

If the merger is still reflected in written language, we should expect to see a higher ratio of *sin/sit* in areas that do not make the distinction (i.e., in the western part of Denmark), because it is used in all contexts.
Overall, we find 420 locations that contain reviews with the possessive pronouns. Note that if the reviews for a location do not include sin/sit, the ratio is 0. If these are the only pronouns recorded for a location, the ratio is 1.

In order to find a regional distribution, we divide our data into NUTS-2 regions. Figure 5 shows the regional distribution according to these regions.

The map shows two distinct parts, at a line drawn at 11 degrees longitude. It approximately divides the country into two even parts, with Sealand (with the capital, Copenhagen) and Bornholm in the east, and Fyn and Jutland in the west. We have 159 data points in the eastern part, 261 in the western. Aggregating over all ratios for each of the two parts, we find a ratio of 0.27 in the west, and of 0.21 in the east. The $\chi^2$ test on the contingency matrix comes out highly significant, i.e., the ratio of sin/sit among the possessive pronouns is highly correlated with regions. This is likely due to the merger in the west, and is in line with the traditional dialectological distinction.

In this case, we were able to confirm a dialectological distinction with our data-driven approach.

### 7.2 Swear words across location, gender, and age

Traditionally, dialectal language is tightly linked with the region where it is natively spoken, to the degree that it is rarely heard (or understood) outside that region. One way this happens is when speakers of a dialect change to the standard language when addressing non-dialect speakers. On the internet, however, it is not always obvious whether the audience is local or global. The question is whether people still use dialect when they are writing reviews, which will be publicly available on the internet.

We consider the word *træls*, which is primarily found in the Jutland dialect. Used as an adjective, it suggests that some situation or state of affairs is unpleasant or unwanted. Interestingly, the word has no exact translation equivalent in standard language, which might make it harder for people to abandon the use of the word as they move outside their home region, virtually or physically.

Figure 6 maps out the regional distribution of the word following NUTS-2 regions. The map shows that the word is mostly found in the western part of the country. Of the 841 reviews with at least one occurrence of *træls*, 84% were written by people living west of the 11 degrees longitude line dividing Sealand and Funen. The word is also used by people living in Copenhagen; the number of users there, however, is small compared to the population of the city. These are most likely Jutlandic citizens living in Copenhagen.

Furthermore, *træls* has a gender and age distribution which deviates from the population as a whole. Notably, as can be seen in Figure 7, it is much more prevalent in the younger age groups, and the rate of usage falls sharply off as people get older. This might be explained by the fact that the word does not belong to the standard language, and as people grow older, they tend to use more conservative language.

Finally, for comparison, we plot the age and gender distribution of the word *lort*. *Lort* is a curse word (literally: *feces*), although not strong enough to be considered offensive in most contexts, and is frequently used to characterize the same types of situations as *træls*. Here, too, we see that the principal users of the word are young people, although women use this stronger version less than the men of the same age group. This is in line with the findings of [31] that the use of swear words/expletives reduces with age. However, contrary to her study, we do find gender differences, with women using the “softer” (and more dialectal) version more than men.

### 7.3 Compounds and multi-word expressions

In Danish, a considerable part of lexicon entries are multi-word units, such as the words for “today” (*i dag*), “yesterday” (*i går*), or “furthermore” (*i øvrigt*). These expressions have traditionally been written apart, i.e., with an intervening white space between the two words. However, in daily use, these expressions are no longer identified by the speakers as composed of two elements, but grammaticalized, i.e., they are interpreted as a single linguistic unit.

Compound words, on the other hand, are officially always written in one word, i.e., without white space. They are often made up on the spot and thus usually analyzable as two elements. Compounds are a productive source of new words in Danish, and potentially unlimited. In English, in contrast, compounds are almost always written with white
space between the nouns, and this seems to influence how young Danes write compounds. To complicate matters, the official spelling has been reformed within the last 50 years, and many words that used to be written together are now written apart.

As a result, there is a multitude of different spellings in Danish for both compounds and multiword adverbial expressions. We hypothesize that there are age-related differences in the preference for one the various spelling variants. Specifically, we expect younger Danes to split noun compounds more often than older ones, due to the influence of English, while grammaticizing multiword adverbial expressions such as i dag. We expect older language users to maintain the separate writing for adverbials and to not split noun compounds.

In order to analyze this, we collect all combinations of two words (bigrams) that occur more than 10 times in our corpus, and record for each age how often they are written apart or together. We exclude all bigrams that contain the article suffixes -en or -et, as well as all verb forms of “to be”, and function words. This filters out some false positives (such as *waren* (“the goods”) vs. *war en* (“was a”)).

For the resulting list, we compute the ratio of the two-word spelling for each expression and analyze its correlation with respect to age via Spearman correlation. Table 8 shows the Spearman correlation of the separate spelling with age for some of the results, together with the overall distribution of the two spellings in the language.

We find that the separate spelling variant of *i dag* (“today”) and *i går* (“yesterday”) are strongly correlated with age, i.e., older speakers prefer to write it in one word. However, surprisingly, for *i øvrigt* (furthermore), we see a slight anti-correlation (Spearman $\rho = -0.52$). This means that older speakers tend to write it in one single word. This goes against our expectations. Comparison to a dictionary from 1957 [1], however, reveals that these were indeed the prescribed spellings, with *i øvrigt* written in one word. Older speakers thus seem to continue to use prescriptive writing norms acquired in their youths.

For compounds, we find that the most distinct cases involve the extremely productive adjectival prefix super-, either as an adjective intensifier, or as an adjectival modifier of a noun. Interestingly, older speakers also prefer to use the one-word spelling, while younger speakers tend to prefer the separate spelling. There are three examples in our top 20, and in all of them the separate spelling is strongly anti-correlated with age (see Table 8).

In German, where compounds famously abound, words can be combined with a hyphen or simply by concatenation. However, while there are some words that show variation, we do not find any significant correlations with age or gender. This seems to indicate that in German, compound spelling is not as free as in Danish.

### 8 German

#### 8.1 Spelling reform “ss” vs. “ß”

Germany underwent a spelling reform in 1996, with the main goal to simplify spelling. After a transition period, the new spelling was legally binding for schools from 2007 on. One of the most pervasive (and contentious) changes was the replacement of ß with ss after short vowels. Two of the most frequent words affected by this reform were the conjunction *dass/daß*, “that”, and the modal *müssen/müßen*, “to must”.

Younger people (born after 1980) were thus predominantly taught the new spelling in school, so we expect to see predominant use of *dass/müssen*, with occasional uses of the old form (depending on at which age the new spelling was acquired), while older speakers retain the traditional spelling they acquired in their youth to a much greater extent.

Again, we collect frequency counts for the two variants of the conjunction, and expect to see a correlation of the traditional spelling (with ß) with older users.

Indeed, when we run significance tests, we see that usage of the old spelling (*daß/müssen*) is significantly correlated with age (Spearman $\rho = 0.84$). Most occurrences of the old spelling crop up for users older than 30, i.e., people who started school in 1990 and encountered the new spelling only after they had left elementary school (when most of spelling is learned). If we divide the data into two bins, the most significant difference between the usage in the two groups occurs when splitting at the age of 40. We do note, though, that use of the old form is lower than use of the new form for all ages.

Our findings indicate that while the spelling reform has been largely accepted, there is a residual use of the old forms, significantly correlated with age. We do not find a gender effect for this spelling variation.

### 9 Conclusion

Traditional sociolinguistic studies often lack statistical power to draw valid conclusions, while big-data approaches to language studies mostly lack extra-linguistic information that would enable sociolinguistic studies.

In this paper, we presented user-review sites as a possible solution to this dilemma. The data provides a combination of non-canonical textual information with meta-information about the authors, including age, gender, and location, as well as time-stamps.

We presented several pilot studies that show how linguistic features are highly correlated with these socio-economic

<table>
<thead>
<tr>
<th></th>
<th>translation</th>
<th>% sep.</th>
<th>% joint</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>selv om</td>
<td>even though</td>
<td>0.28</td>
<td>0.72</td>
<td>0.98</td>
</tr>
<tr>
<td>i går</td>
<td>yesterday</td>
<td>0.61</td>
<td>0.39</td>
<td>0.82</td>
</tr>
<tr>
<td>ud over</td>
<td>apart from</td>
<td>0.47</td>
<td>0.53</td>
<td>0.79</td>
</tr>
<tr>
<td>i dag</td>
<td>today</td>
<td>0.67</td>
<td>0.33</td>
<td>0.78</td>
</tr>
<tr>
<td>i orden</td>
<td>alright</td>
<td>0.84</td>
<td>0.16</td>
<td>0.74</td>
</tr>
<tr>
<td>i øvrigt</td>
<td>furthermore</td>
<td>0.66</td>
<td>0.34</td>
<td>-0.52</td>
</tr>
<tr>
<td>inden for</td>
<td>inside</td>
<td>0.54</td>
<td>0.46</td>
<td>-0.54</td>
</tr>
<tr>
<td>hjemme side</td>
<td>home page</td>
<td>0.03</td>
<td>0.97</td>
<td>-0.59</td>
</tr>
<tr>
<td>ind til</td>
<td>into</td>
<td>0.38</td>
<td>0.62</td>
<td>-0.70</td>
</tr>
<tr>
<td>lyn hurtig</td>
<td>lightening fast</td>
<td>0.17</td>
<td>0.83</td>
<td>-0.77</td>
</tr>
<tr>
<td>super service</td>
<td>super service</td>
<td>0.97</td>
<td>0.03</td>
<td>-0.79</td>
</tr>
<tr>
<td>i det</td>
<td>in that</td>
<td>0.73</td>
<td>0.27</td>
<td>-0.86</td>
</tr>
<tr>
<td>super hurtig</td>
<td>super fast</td>
<td>0.94</td>
<td>0.06</td>
<td>-0.88</td>
</tr>
<tr>
<td>super god</td>
<td>super good</td>
<td>0.95</td>
<td>0.05</td>
<td>-0.90</td>
</tr>
</tbody>
</table>
variables. This includes the use of emoticons (with and without nose), detection of gender-specific words, and rating behavior, as well as phenomena in Danish (reflexive pronouns, swear words, and compounding) and German (spelling reform).

The focus of this work has been to evaluate the suitability of the resource. Given the robustness of our findings, we plan on investigating the consequences of language variation among groups for NLP tools. More concretely, do NLP tools perform equally well for all demographic groups, and if not, can we use the findings of our study to improve them?

We also plan to a) relate the data to extra-linguistic information and b) augment it with information on the grammatical information.

10 Acknowledgements

This research is funded by the ERC Starting Grant LOW- LANDS No. 313695. The authors would like to thank the anonymous reviewers for their helpful suggestions.

11 References