Impoverished Language Technology: The Lack of (Social) Class in NLP

Amanda Cercas Curry, Zeerak Talat, Dirk Hovy
Bocconi University, Mohamed Bin Zayed University of Artificial Intelligence, Bocconi University
amanda.cercas@unibocconi.it, z@zeerak.org, dirk.hovy@unibocconi.it

Abstract

Since Labov's (1964) foundational work on the social stratification of language, linguistics has dedicated concerted efforts towards understanding the relationships between socio-demographic factors and language production and perception. Despite the large body of evidence identifying significant relationships between socio-demographic factors and language production, relatively few of these factors have been investigated in the context of NLP technology. While age and gender are well covered, Labov's initial target, socio-economic class, is largely absent. We survey the existing Natural Language Processing (NLP) literature and find that only 21 papers go beyond merely mentioning socio-economic strata in passing. However, the majority of those papers do not engage with class beyond collecting information of annotator-demographics. Given this research lacuna, we provide a definition of class that can be operationalised by NLP researchers, and argue for including socio-economic class in future language technologies.

Keywords: Social Class, Marginalisation, Linguistic Diversity

1. Introduction

A salient aspect of identity formation is the creation of class, or socio-economic identity. Certain accents, phrases, and constructions are considered either upper, middle, or lower class, or a host of other references to socio-economic status.

Labov (1964) was the first to systematically investigate this relation. He observed that New Yorkers with higher socio-economic status tended to pronounced the /R/ sound after vowels, whereas the traditional dialect dropped it. He devised a study to quantify this observation by asking clerks in various department stores (a proxy for socio-economic status) for items found on the fourth floor, then recorded how many of Rs were dropped. He found a clear anti-correlation between the socio-economic status of the store (and hence presumably the speakers there) and the number of dropped Rs: the higher the status, the fewer dropped Rs.

Since Labov's 1964 intervention on social stratification and language, linguistics has dedicated concerted efforts towards understanding how different socio-demographic factors influence the production and perception of language, and how speakers use them to create identity (Eckert, 2012). Despite the large body of evidence showing the relationships between language and demographic factors (the "first wave" of socio-linguistic variation studies), relatively few socio-demographic factors have been investigated in the context of natural language processing (NLP) technology.

Existing work on socio-demographic factors has predominantly focused on how much a certain linguistic variable signals age, ethnicity, regional origin, and gender (Johannsen et al., 2015). Almost none of the NLP works have engaged with the second and third wave of sociolinguistics, i.e., how variation creates local identity and drives language change. To address this gap, we focus on socio-economic status in NLP.

Filling this lacuna could provide useful future research avenues for computational sociolinguistics and social science, which often depends on the proper stratification of data into socio-demographic categories. More broadly, excluding a crucial sociodemographic factor like social class from consideration impoverishes NLP's capability to counteract social biases in its tools and datasets.

Contributions We document the lack of NLP work dealing with socio-economic status. We survey how socio-economic status is measured in the NLP literature and contrast this with metrics used in the social sciences. We conclude with some recommendations for future research.

2. What is Social Class?

Social stratification refers to the grouping of people according to socio-economic status (SES) based on factors like income, education, wealth, and other characteristics, with different groups being distinguished in terms of power and prestige (Saunders, 1990). There are different systems of social stratification, including the Indian caste system, clans or tribes, and the Western hierarchical class system.

Exactly how many social strata there are is unknown and likely varies from region to region. However, in Western cultures it is common to see at least three strata referring to upper, middle and

LREC-COLING 2024, pages 8675–8682

lower class people. Other systems refer to blue collar and white-collar jobs. Recently, the Great British Class Survey (GBCS, Savage et al., 2013) has taken an empirical approach to understanding the different social strata, and they propose a seven-level system for the United Kingdom. Their stratification is based on economic, social and cultural capital: elite, established middle class, technical middle class, new affluent workers, traditional working class, emergent service workers, and precariat. The derive these classes from a survey conducted over British citizens that received more than 160K responses.

3. The Impact of Social Class on Language

Social class shapes people's everyday experiences by granting or limiting access to resources. Social stratification has a significant impact on people beyond power and prestige, for instance, lower socio-economic status has been linked to worse health outcomes and higher mortality (Saydah et al., 2013). Socio-economic status also influences language: Labov's germinal work showed there are differences in pronunciation, Bernstein (1960) showed that children from working class families had significantly smaller vocabularies even when general IQ was controlled for, and Flekova et al. (2016) has recently shown that there is significant lexical and stylistic variation between social strata.

Socio-economic status affects language use from the very early stages of development. Bernstein (1960) posits that language takes on a different role in middle- and working-class families, where middle-class parents encourage language learning to describe more abstract thinking. In working-class families, parents are *limited* to more concrete and descriptive concepts. Parents from lower SES tend to interact less with their children, with fewer open-ended questions than parents from higher SES, which shapes language development (Clark and Casillas, 2015). While Usategui Basozábal et al. (1992) shows that education reduces this language gap, one's level of education has traditionally been one of the key factors in determining social class and potential for upward social mobility. Moreover, a person's accent is a strong marker of SES, reported to lead to anxiety and discrimination (Levon et al., 2022).

Given the well-documented effects that socio-economic status has on language development and use, it stands to reason that NLP should carefully consider social class as a variable.

4. Measuring socio-economic status

As social class encompasses more than a single factor, e.g., income, measuring and classifying people according to socio-economic status is a non-trivial task. Social class may be measured objectively through measures of socio-economic status (SES) or subjectively, by asking participants to self-report.

Objective In terms of objective measures of social class, education, income, and occupation are the most widely used factors to measure social class (Kraus and Stephens, 2012). Education affords access to higher salaries, more prestigious occupations and higher levels of cultural capital. Income is the most direct measure of an individual's access to material goods and services. Finally, occupation is a strong indicator of prestige and other formative experiences.

More recently, the GBCS (Savage et al., 2013) asked participants about their economic, social and cultural capital based on a framework described by Bourdieu (2018). Economic capital refers to one's income and assets[1], *social capital* is measured in terms of the prestige of those in one's social circle[2] and *cultural capital* in terms of the type of cultural activities in which they participate (e.g. attending the theatre vs football matches). Based on the results from the survey, Savage et al. (2013) propose a seven-level stratification extracted from latent class analysis on 160K participants from the U.K. This fine-grained classification shows societal changes and fragmentation in the middle-class.

Subjective In terms of subjective measurements, social class is how much one believes they have relative to others. People's perception of where they stand in terms of social class has important psychological effect even when controlling for objective measures, supporting the idea that subjective class is an important measure. The general recommendation is to use the Macarthur scale (Adler et al., 2000), where people are asked to place themselves on a ladder, with higher levels representing those who are more privileged.

Because there may be discrepancies between one's subjective and objective SES, the American Psychological Association (APA)[3] recommends measuring a participant's level of education, income, occupation, and family size and relationships, as well as subjective social status (Diemer

[1] Namely, whether one owns their house or rents it and the amount of savings.

[2] As determined by their status scores according to the Cambridge Social Interaction and Stratification (CAMSIS) scale.

[3] APA: Measuring SES

et al., 2013). These recommendations add a layer of difficulty for researchers because (1) the questions are intrusive (for example in many countries it is considered impolite to discuss salaries) and refer to sensitive topics, and (2) while gender and race are a single data point, accurately measuring socio-economic status requires a minimum of four questions that need to be aggregated.

For these reasons, the majority of NLP work has focused on one particular aspect of social class as a proxy. For example, Lampos et al. (2014) use occupation only as a proxy and Flekova et al. (2016) use only income. However, income is only one small part of social class and does not capture its nuances and the over-emphasis on income and occupation of most social classification systems has been criticised by some feminist scholars for how it interacts with some social norms, e.g., the expectation of women to be homemakers (Skeggs, 1997; Crompton, 2008).

5. Survey of NLP Literature on Social Class

Here, we analyse existing literature in NLP that deal with SES in some way. As a first step, we collect the bibliography file for the ACL anthology and search for the occurrence of terms in the titles and abstracts.[4] We also experimented with terms like 'occupation' and 'education' which may be used a proxy of social class, however we find that these papers focus on either gender bias (in the case of occupation), or education in and of itself without engaging with social class in any way. In addition, we refer back to feminist scholars who criticise metrics such as income and occupation as individual class markers in patriarchal societies (see Skeggs (1997); Crompton (2008)). Our initial search yielded 78 papers, however, many of them do not engage with the topic at all and only mention social class in passing, e.g. 'communication between people from different religion, caste, creed, cultural and psychological backgrounds has become more direct'. After removing such papers, only 21 papers remain.

Of these 21 papers, four focus on NLP in low- and middle-income countries/regions for social equity but do not directly model language. Three papers collect SES-related metrics, but do not use this information in their analyses. Finally, the remaining papers deal directly with modelling language according to SES (e.g. predicting the income of Twitter users).

Although (as expected) the majority of papers are dealing with English, we find a wide variety of languages given the small sample (English, Danish, French, Hindi, Marathi, Russian).

5.1. How NLP Measures SES

So far, no systemic method to measure SES has been proposed for NLP. Table 1 shows the full list of papers and their SES measurement. The majority focus on one or two aspects of SES, such as income (9) and education (5). Only five papers refer explicitly to a class system, using a two- or three-level classification, however none use objective measures of class. One paper uses restaurant prices as a proxy for socio-economic status. In addition, in terms of granularity, many of these studies do not collect data on an individual level but rather use reported statistics for a given country/area using census data.

We conclude from the survey that socio-economic status is rarely reported on in NLP literature, where most data is collected from urban citizens and university students, or from middle- to upper-class sources like news outlets. Low-SES is only specifically collected to cover this subset of the population for a given study, possibly due to the increased difficulty in accessing people from less privileged backgrounds when research is done in universities and affluent urban areas. However, sourcing data from lower-SES participants still affords quality data while also offering supplemental income (Abraham et al., 2020). As NLP technologies are becoming increasingly ubiquitous in society, we should endeavour to include all subpopulations to ensure fair and equitable technologies. In addition, NLP also serves a role as cultural anthropology and should reflect the reality of language use across populations.

6. Measuring Class for NLP Practitioners

A standout finding from the survey is the lack of a unified measurement of socio-economic status in the literature. This makes it difficult to compare language varieties across studies and datasets or to get a clear picture of whose language exactly NLP research is mapping. In addition, it remains unclear how some of the proxy metrics used (such as geolocation or income) relate to class and language variety. Section 4 provided an overview of possible metrics for socio-economic status. Based on this, we make some recommendations for the NLP community:

- Where participants are explicitly recruited for a study (i.e. the data is not collected en masse from social media), researchers should try to take objective measures of the participants' SES (see 4).

[4] The full list of our search terms: 'social class', 'caste', and 'socio-economic', 'income', 'education', 'occupation', 'white/blue collar', 'upper/middle/lower class'

	Measurement	**Granularity**
Lampos et al. (2014)	Unemployment	Country
Preoţiuc-Pietro et al. (2015)	Occupation	Individual
Flekova et al. (2016)	Income	Individial
Hasanuzzaman et al. (2017)	Income	Individual
Giorgi et al. (2018)	Income, Education	County (census data)
Zamani et al. (2018)	Income, Education, Unemployment	Country-wise
Degaetano-Ortlieb (2018)	Class (high, low)	Individual
Van et al. (2019)	Income, poverty education	State-level (census)
Jawahar and Seddah (2019)	Income, geolocation	Neighbourhood-level
Basile et al. (2019)	Restaurant price	Individual
Ghazouani et al. (2019)	socio-economic status	Mixed
Abraham et al. (2020)	Income, area	Group
Tafreshi et al. (2021)	Income, education	Individual
Abbasi et al. (2021)	Income, education	Individual
Strømberg-Derczynski et al. (2021)	SES (high, mix, unknown)	Aggregated by dataset
van Boven et al. (2022)	Low-income countries	Country
Ngao et al. (2022)	Low-income countries	Country
Grützner-Zahn and Rehm (2022)	GDP	Country
Cole (2022)	Class (high, low)	Individual
Malik et al. (2022)	Caste, occupation	General (bias)
Hržica et al. (2022)	Class (middle)	Group

Table 1: Papers included in the survey along with the metric used to assess socio-economic status and the granularity of the metric is.

- Alternatively, aim to report at least their subjective SES[5] following the Macarthur scale, which has already been scientifically validated.

- Data collected from social media rarely contains such information at the individual level – while people may list their occupation in their biographies, these tend to be biased towards high-prestige occupations (Guo et al., 2024), leaving lower status occupations in the dark. Researchers should endeavour to contextualise the data collected by considering and appropriately describing the general statistics of the social media platform used (e.g. Ghazouani et al. (2019) aim to asses the socioeconomic status of X – formerly known as Twitter – users).

- If all else fails, report socio-economic status in whichever way is possible (for example by using some of the proxy metrics discussed in this paper, or e.g. Cercas Curry et al. (2021) use level of education and the university's prestige, though these may be dependent on culture).

Properly documented and contextualised data collection is crucial for more equitable NLP (and more broadly linguistics) research.

7. Related Work

While there has been an uptick in the number of papers tackling gender and racial bias in NLP, work considering other under-privileged communities has lagged behind. In a survey of Computational Sociolinguistics, Nguyen et al. (2016) point to the lack of self-reported explicit labels in online user profiles as possible cause, with work focusing on occupation (often mapped to income or other variables.

As a way to mitigate and document biases in NLP, Bender and Friedman (2018) ask for the socio-economic status of both the speakers and the annotators to be declared, however they do not suggest any standardised way to measure or report this. Field et al. (2021) (from whom we borrow our methodology) conduct a survey focus on race but also call for more diversity in NLP in terms of the broader inclusion other underprivileged people such as those from lower socio-economic status.

8. Directions for Future Research

NLP has two main purposes: (1) as a descriptive tool for current language use, and (2) as a service for everyday life. In both cases, NLP must be able to serve and represent people equally. With the ubiquity of language technologies, it is no longer the case that only the wealthiest need to access these but by excluding those less privileged from NLP datasets we are crippling our technologies in

[5] As this is a one-point measure and is not as intrusive as concrete questions about income.

their ability to serve all of humanity. NLP systems are enforcing a standard of language by limiting the lects they represent.

Future research should establish how the current and proposed metrics and models represent language varieties. Some questions for future work are whether subjective class is more predictive of language use than objective class, how fine-grained of a classification is needed, and how socio-economic status interacts with other socio-demographic factors like age, gender, race or region when it comes to language.

Creating new datasets and tools to identify social class distinctions in text would not only help build fairer NLP technology, but also benefit related disciplines that use NLP tools to stratify their data along socio-demographic lines.

9. Conclusion

We have explored the definition of social class, how it can be measured and its effect on language use. We then surveyed existing work in the ACL anthology dealing with socio-economic status and its components and found significant gaps. First, very little work considers social class despite its well-documented effect on language, and second, there is currently no systematic way to measure socioeconomic status in the few papers that do report it. We encourage researchers to ensure diversity in collected datasets and to be more diligent in reporting accurate socio-demographics for their participants.

Limitations

Our survey has provided an overview of work in NLP engaging with socioeconomic status, however, we focused only on papers included in the ACL anthology. This means some work that has been published in other venues has been excluded. In addition, our keyword search is done only on titles and abstracts, though any work meaningfully engaging with SES is likely to mention it. Finally, our survey and suggestions have focused mainly on Western definitions of SES and social class that may not be applicable in other cultures.

Acknowledgements

Amanda Cercas Curry and Dirk Hovy were supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). They are members of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

10. Bibliographical References

Ahmed Abbasi, David Dobolyi, John P. Lalor, Richard G. Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3748–3758, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826, Marseille, France. European Language Resources Association.

Nancy E Adler, Elissa S Epel, Grace Castellazzo, and Jeannette R Ickovics. 2000. Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, white women. *Health psychology*, 19(6):586.

Angelo Basile, Albert Gatt, and Malvina Nissim. 2019. You write like you eat: Stylistic variation as a predictor of social stratification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2583–2593, Florence, Italy. Association for Computational Linguistics.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Basil Bernstein. 1960. Language and social class. *The British journal of sociology*, 11(3):271–276.

Pierre Bourdieu. 2018. Distinction: a social critique of the judgement of taste. In *Inequality Classic Readings in Race, Class, and Gender*, pages 287–318. Routledge.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eve V Clark and Marisa Casillas. 2015. First language acquisition. In *The Routledge handbook of linguistics*, pages 311–328. Routledge.

Amanda Cole. 2022. Crowdsourced participants' accuracy at identifying the social class of speakers from South East England. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pages 38–45, Marseille, France. European Language Resources Association.

Rosemary Crompton. 2008. Class and stratification. *Journal of Social Policy*, 38:361–362.

Stefania Degaetano-Ortlieb. 2018. Stylistic variation over 200 years of court proceedings according to gender and social class. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 1–10, New Orleans. Association for Computational Linguistics.

Matthew A Diemer, Rashmita S Mistry, Martha E Wadsworth, Irene López, and Faye Reimers. 2013. Best practices in conceptualizing and measuring social class in psychological research. *Analyses of Social Issues and Public Policy*, 13(1):77–113.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41(1):87–100.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

Dhouha Ghazouani, Luigi Lancieri, Habib Ounelli, and Chaker Jebari. 2019. Assessing socioeconomic status of Twitter users: A survey. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 388–398, Varna, Bulgaria. INCOMA Ltd.

Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H. Andrew Schwartz. 2018. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1172, Brussels, Belgium. Association for Computational Linguistics.

Annika Grützner-Zahn and Georg Rehm. 2022. Introducing the digital language equality metric: Contextual factors. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 13–26, Marseille, France. European Language Resources Association.

X. Guo, D. Handzlik, J. Jones, , and S. Skiena. 2024. The evolution of occupational identity in twitter biographies. In *Proceedings of the Int. AAAI Conf. on Weblogs and Social Media (ICWSM 2024). Forthcoming*.

Mohammed Hasanuzzaman, Sabyasachi Kamila, Mandeep Kaur, Sriparna Saha, and Asif Ekbal. 2017. Temporal orientation of tweets for predicting income of users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 659–665, Vancouver, Canada. Association for Computational Linguistics.

Gordana Hržica, Chaya Liebeskind, Kristina Š. Despot, Olga Dontcheva-Navratilova, Laura Kamandulytė-Merfeldienė, Sara Košutar, Matea Kramarić, and Giedrė Valūnaitė Oleškevičienė. 2022. Morphological complexity of children narratives in eight languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4729–4738, Marseille, France. European Language Resources Association.

Ganesh Jawahar and Djamé Seddah. 2019. Contextualized diachronic word representations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 35–47, Florence, Italy. Association for Computational Linguistics.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China. Association for Computational Linguistics.

Michael W Kraus and Nicole M Stephens. 2012. A road map for an emerging psychology of social class. *Social and Personality Psychology Compass*, 6(9):642–656.

William Labov. 1964. *The social stratification of English in New York city*. Ph.D. thesis, Columbia University.

Vasileios Lampos, Daniel Preoţiuc-Pietro, Sina Samangooei, Douwe Gelling, and Trevor Cohn. 2014. Extracting socioeconomic patterns from the news: Modelling text and outlet importance jointly. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 13–17, Baltimore, MD, USA. Association for Computational Linguistics.

Erez Levon, Devyani Sharma, and Christian Ilbury. 2022. Speaking up: Accents and social mobility. Technical report, The Sutton Trust.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Narshion Ngao, Zeyu Wang, Lawrence Nderu, Tobias Mwalili, Tal August, and Keshet Ronen. 2022. Detecting urgency in multilingual medical SMS in Kenya. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 68–75, Online. Association for Computational Linguistics.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, Beijing, China. Association for Computational Linguistics.

Peter Saunders. 1990. *Social class and stratification*. Routledge.

Mike Savage, Fiona Devine, Niall Cunningham, Mark Taylor, Yaojun Li, Johs Hjellbrekke, Brigitte Le Roux, Sam Friedman, and Andrew Miles. 2013. A new model of social class? Findings from the BBC's Great British Class Survey experiment. *Sociology*, 47(2):219–250.

Sharon H. Saydah, Giuseppina Imperatore, and Gloria L. Beckles. 2013. Socioeconomic status and mortality. *Diabetes Care*, 36(1):49–55.

Bev Skeggs. 1997. *Formations of class & gender: Becoming Respectable*. Sage Publications Ltd.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

Elisa Usategui Basozábal et al. 1992. La sociolingüística de Basil Bernstein y sus implicaciones en el ámbito escolar. *Revista de educación*.

Hoang Van, Ahmad Musa, Hang Chen, Stephen Kobourov, and Mihai Surdeanu. 2019. What does the language of foods say about us? In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 87–96, Hong Kong. Association for Computational Linguistics.

Goya van Boven, Stephanie Hirmer, and Costanza Conforti. 2022. At the intersection of NLP and sustainable development: Exploring the impact of demographic-aware text representations in modeling value on a corpus of interviews. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2007–2021, Marseille, France. European Language Resources Association.

Mohammadzaman Zamani, H. Andrew Schwartz, Veronica Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. 2018. Residualized factor adaptation for community social media prediction tasks. In *Proceedings of the 2018 Conference* *on Empirical Methods in Natural Language Processing*, pages 3560–3569, Brussels, Belgium. Association for Computational Linguistics.