## Word embeddings

Word embeddings are continuous, dense high-dimensional vector spaces that can approximate the semantic relations between words. Initially used for solving NLP problems of analogies and word pair similarity, they quickly became appropriated for use in research in bias. They were found to accurately capture racial, ethnic and gender biases that were present in the texts they were trained on(Brunet et al., 2018; Caliskan et al., 2017; Lewis & Lupyan, 2020)¤copied - shift order and add. With that, they promise the possibility of automatically recognising hate speech, and potential avenues to correction ¤CIT (DEBIASING WORD EMBEDDINGS PAPER). But these same biases can be used to infer the culture that words are produced in. That is, culturally "boxing" may be stereotyped as a male sport and horseriding a female sport, but it may also be considered that culturally we associate them as such. They provide, as such a means by which to examine culture as an inherently biased, heuristic or schematic endeavour, all words being suffused with such assocations.

This is a direction subsequent research has taken and applied in various domains of cultural research. Garg et. Al (2018) used Google Ngrams word embeddings to capture gender and ethnic biases of occupations through the twentieth century. Kozlowski et al. (2019) used Google Ngrams word embeddings to capture class biases in language and analysed how different dimensions of class related to one another through the twentieth century. More recently, Arseniev-Koehler and Foster (2020) trained word embeddings on a corpus of over one hundred thousand New York Times articles and were able to find significant biases regarding obesity and key cultural concepts such as socioeconomic status, gender, morality and health. Laura Nelson used word embeddings trained on abolitionist texts of the 19th century to map biases of an intersectional nature - towards black and white men and women – verifying these all in a close reading of the texts in question. I will focus on the studies by Garg et al. and Kozlowski et al., and describe how I propose to synthesize these approaches to investigate class in language from a new angle.

Garg et. Al’s approach used word embeddings trained on Google Ngrams by Hamilton et al. (2016) to map out gender and ethnic biases towards occupations in the US labour force. Training a regression algorithm on these projections, they were able to predict the actual gender and ethnic compositions of occupations to a high degree throughout the 20th century. They found that they could reproduce the discrete shifts towards the more balanced composition of the labour force after successive waves of feminism (as measured by a decrease in correlation between word embeddings occupation biases for the corresponding decades). They were also able to depict the prevailing stereotypes of each gender and ethnicity by listing the adjectives (within various domains such as professional competency, appearance, etc) most associated with an aggregate of the 5,000 most common names associated with the gender or ethnicity. The top 3 words associated with Asian Americans in 1910, for example, were “irresponsible”, “envious” and “barbaric”, but by 1990 this had shifted to “inhibited”, “passive”, “dissolute”. This demonstrated that word embeddings could be used to capture objective social relations as well as prevailing stereotypes, by virtue of the biases they incorporate.

The brilliant Geometry of Culture study (Kozlowski, 2019) extends this analysis to the study of class. This study uses word embeddings to consider change in the cultural perception of class in the 20th century. It considers 5 different dimensions of class according to leading theories in the domain - affluence, cultivation, ownership of capital vs. employment, education and status (as well as race, gender and morality) - and analyses the relationship between them over time. It does this mainly by analyzing the similarity between dimensions (by virtue of cosine similarity), finding a relative stability between them throughout the 20th century, with the exception of education which becomes more closely related to affluence (following the great expansion of the education sector as a means for social mobility and distinction). Likewise, it established relative levels of drift of words with respect to each class dimension over time - words becoming, in general, drifting from their meanings over time. It validated these dimensions of culture by producing significant correlations between the cultural dimensions they created in their word embeddings space and i) an Amazon Mechanical Turk survey they fielded, which studied respondent ratings of a set of words on various scales of class, as well as ii) a large semantic differential study performed by Jenkins et al. in 1959 (An atlas of semantic profiles) where respondents rated words on 20 different scales (which the researchers took to be core dimensions of semantics within language).

These studies complement each other intimately, but it is in their respective lacks that I think an important synthesis lies. Garg concentrated on validating biases in word embeddings by showing their mirroring of objective structures in society. But it is doubtful whether this means that word embeddings capture “biases” – it could just as well by a consequence of language discussing the world outside. Biases here may just be frequent co-occurrence of words that follow frequent co-occurrence of objects – female bias towards domestic worker occupations may just be due to a higher frequency of female domestic workers to talk about. Kozlowski, on the other hand, validated biases by showing how word embeddings word projections mirror human ratings of words – and thus validates them from an (inter)subjective perspective (in contrast to Garg’s objective approach). Though these are closer to validating that they capture biases, they i) face the problem that the human ratings may not be biases but considerations of their associations from an objective reality (cf. the previous example of domestic workers) – i.e. their distance from objective reality is not measured - and ii) they do not go in deep enough detail of detail of what the notion of class is that they are actually capturing with the word embeddings approach. Regarding the first point, I believe it is in the *discrepancy* between objective and subjective approaches of validation, that a meaningful concept of bias should be measured. I will return to this in the methodology section. Regarding the second point, I believe that what word embeddings can most meaningfully capture is the *perceived homology between the space of lifestyles and the structure of the economy.* That is, our intuitive sense of the “classedness” of things, their relative position in society – how we intuit the kind of things that rich people do, poor people drink, bourgeois people talk about, middle class people dress in. The language of “class”, antiquated as it feels, does not matter – it is about the sense of position of things. To unpack this, I will turn to Bourdieu, who provides the deepest theoretical explanation of it – considering his notion of “homology” first, followed by his notion of “class sense”.

### Structural homology of the social space

A central thesis of Bourdieu’s magnum opus, Distinction, is the structural homology of the social space. The intuition is simple: the space of culture is closely related to the structure of the economy. Taste in art, music or food, participation in sport, manners of speaking, gestures of the body – all of these align with an individual’s position in the economy. The direction of determination is not one-way, but the alignment is a one-to-one correspondence, or, “homology”. Individuals tend to engage with cultural practices that are appropriate to their position.

This homology means that both lifestyles and the economic division of the labour force can be modelled on the ordering metaphor of “space”. In Distinction, he seeks to map out “the space of lifestyles" (871) of France in the 1960's and how its accordance with the structure of labour. He does so by conducted a survey inquiring into various aspects of cultural engagement - taste in art and music, participation in sports, food, clothing and manners – and demographic background information on a large crossection of France (n = ?). Applying a novel technique in geometrical data analysis, multiple correspondence analysis he was able to visualize each of these domains of culture, as well as clusters of occupations, on a 3-dimensional space – as can be seen in figures 1,2,3.

The space was mapped out according to 3 parameters: "volume of capital”, "composition of capital", and to a lesser degree “trajectory of capital”. The volume of capital is as simple as it sounds - it is the amount of capital an individual has to their name, which aligns with our intuitions of the rich and poor but diverges with the Marxist conception of owner of and worker for capital. He later describes it as volume of accumulated labour, which can be exchanged or used to appropriate other people’s labour – in short, "pre-emptive rights over the future" (savage et al. 2013). The composition of capital is his original contribution here – it is a measure of the ratio between economic and cultural capital of an individual's capital. Cultural capital is a multifarious concept but at its core is a learned set of dispositions that manifest in one’s “taste”, as well as the capacity to create products in accordance with it. It is later defined in terms of three subcategories: "institutional" cultural capital (validation from institutions like the university), "objectified" (material cultural objects like works of art) and "embodied" (habits and capacities of the mind) capital – though the latter is the most integral to his thinking(243 forms). He also considers a third form of capital later on, social, (248 forms get some quotes!!)which refers to the power one has through one's network - the people one knows, the amount of them and their collective capacity. This latter one is downplayed as more relevant to the analysis of communistic society, but is still relevant. Each of these forms of capital can be exchanged for one another, and thus are theoretically integrated in the social space.

Classes can be distinguished in this space by volume of capital, but there can be radically different compositions of capital within a class - the classic example being "the curate with high status but low income and the bookmaker with low status but high income" (Runciman, 1972, p. 45). (69). Measuring simply on volume of wealth, or degree of ownership of the means of production, would obscure from our intuition that these two examples are much closer to one another than, say, a garbage collector. But it is important to note that Bourdieu refrains from theorizing classes, his stated goal being "not to propound a theory of social classes but rather . . . to uncover principles of differentiation capable of accounting . . . for the largest possible number of observed differences" (Bourdieu, 1990b, p. 117).(138). Rather than the “big classes” that we are familiar it is – it is better to consider it on this metaphor of space, within which certain clusters form, these clusters making possible the formation of distinct class consciousnesses .

With respect to the space of lifestyles, it is important to note how cultural artifacts differ along the "composition of capital" axis: they run from the refined and rare to the vulgar and common, or, the legitimate and the illegitimate. The legitimate taste is where cultural capital concentrates. Here, taste values form over function, style over substance, manner over matter. In food, the rich have a highly ritualised procedure of dinner that emphasises aesthetic appeal, while the poor emphasis ideas of "honesty", "plenty" and "freedom". In clothing, the fashion and formality of dress contrasts with the functional clothing of the poor, where "value for money" and "durability" are the orders of the day. Cultivation has, of course, much in common with affluence; as he remarks, as one moves up the economic spectrum, lifestyles start becoming "the stylization of life". That is, one may make a free choice in what culture one pursues, while the poor are restricted by the necessities of life - "the choice of necessity" (73, atkinson). Tastes follows these contours much of the time, but also differs in important ways. Social classes like the nouveau riche have more in common culturally with the poor - their lifestyle remaining "very close to that of the working classes as regards economic and cultural consumption". The petit bourgeoisie, on the other hand, adopt the legitimate tastes of the bourgeoisie but are accompanied by awkwardness and embarrassment, in comparison to the bourgeoisie's "ease" (SEE THAT STUDY ON A BOARDING SCHOOL AND GCBS¤) of consumption and reflection - knowing that they constitute part of the subjectivity which determines legitimate taste, a native and not a foreigner to its land.

### Class sense

Beyond these notions of the objective dynamics of class relations in economy and culture, Bourdieu also considers the subjective element of it. We have an innate "class sense" (atkinson, 74) that senses these objective relations. Our class sense is the ability to naturally pinpoint in a snapshot where others stand in relationship to us, as well as we to them - "a sense of one's place" (466, distinction); who is top, who is bottom. As Atikinson put it, it is “a relational sense of difference and similarity, of antipathy and sympathy, of ‘one’s place’ and the place of others, and ultimately of distance and proximity in social space, based on the ‘reading’ of the signifiers of symbolic space borne and performed by bodies”. It is something that co-constructs the meaning of cultural objects, from the more obvious things like doing sport, enjoying art and music and having a job, to the "apparently most insignificant techniques of the body — ways of walking or blowing one’s nose, ways of eating or talking" (class/classif). No object is simply itself, it is inherently also a social object, and that aspect is constructed, or sensed, by this capacity within us.

Class sense is an aspect of Bourdieu’s more well known notion of “habitus”. The habitus is most frequently described as the "enduring set of dispositions" that make up our perceptions and actions. It is a cultural structure that shapes many of our beliefs, values and thoughts. Habitus has a dual function: "the capacity to produce classifiable practices and works, and the capacity to differentiate and appreciate these practices and products (taste)" - the dynamic relationship of these being how "the represented social world, i.e. the space of lifestyles, is constituted." (871). It determines both the outcome of our creative practices (understood broadly – down to the level of daily life), as well as how we appreciate the creative practices of others (our taste). This latter component is composed of a mass of "schemes", schemata that classify everything that come its way. These schemata both determine what kinds of things we like, but also allow us to understand where things are in relation to other things. As Bourdieu puts it, they are "a practical mastery of distributions which makes it possible to sense or intuit what is likely (or unlikely) to befall — and therefore to befit — an individual occupying a given position in social space." They classify objects in social space and therewith the individuals that are related to them. This is what the notion of “class sense” refers to, an ability to intuit objects position in the social space mapped out by Bourdieu in his study.

These "schemes of the habitus... function below the level of consciousness and language, beyond the reach of introspective scrutiny or control by the will." (Bourdieu, Marxists.org extract) This class sense thus doesn't have to be voiced in the conceptual language of class, or even voiced at all. Though the language of class was important in Western countries in the 20th century, class sense is just as intimately expressed in terms like rich, poor, good, bad, tasteful, tacky - or more poignantly in slang like "posh", "toff", "chav" or "waster". This class sense is not universally common, each individual only having imperfect access to the whole, nor does it necessarily pertain to the “big classes” of the lower, middle and upper – though these are often good approximations. It is perhaps best elucidated by comedy or humour more generally, which takes upon these underlying schematic understandings and brings them into the light. Making fun of people who are in some way reprehensible or lower, making caricatures which play upon prevalent stereotypes, pejorative names for people - these are all means by which people are made aware of what is socially lower, and require like schemata in the audience to be able to resonate. Lower, of course, is in this case not just determined by capital – but across many dimensions – race, gender, morality, personality – anything that could constitute a “type of guy”.

### Class sense & word embeddings

This notion of class sense is very useful when considering how stratification manifests in language. It implies that there is an inherent sense of where each object fits into the cultural landscape, which both patterns our interpretation of other people and the cultural objects they engage with, as well as our choices of objects that we engage with. This gets reflected in what language we choose to use and how we talk about the world.

This aligns importantly with word embeddings, since word embeddings seek to reproduce implicit biases we have by virtue of the language we use. The promise of word embeddings is the tapping into conscious or unconscious associations we have of cultural phenomena, by means of the algorithmic modelling of the word contexts of their corresponding words. There are certain words and certain kinds of words that fall naturally into the mind when thinking of particular cultural objects. If we are prompted "Trump" we might think president, old, businessman, money, tanning salon, American south and if we are prompted "glasses", we might think of cups, monocles, intelligence or eyes. These associations are things that come unprompted and resemble commonly available stereotypes, though are never quite fully aligned, dependent as they are to our cultural exposure. Pick up on which words commonly appear together, generalise in the frame of a language model, and a schematic understanding of the cultural multivalence of words and their signified phenomena can be built. If this were to be the utterances of one individual, this would hypothetically approximate their "societal sense" (class sense here being a subsection or angle of this sense), but in the case of large corpora, it may be hypothetically approximate a "general social sense" - the aggregated multiplicity of meanings across a society. As Koslowski [sic]¤ put it - the "shared understandings". We can imagine here to be the realm of stereotypes - hence why word embeddings found their first application in the field of bias research.

### Word embeddings complementariness to the notion of space

Likewise, the ordering metaphor of “space” aligns deeply with the capabilities of word embeddings. The concept of space as used by Bourdieu is continuous (as opposed to discrete or qualitative), linear and multidimensional. This aligns exactly with the dimension projecting method of word embeddings. As shown by Kozlowski et al. (2019), can be used to project words onto various dimensions of culture. These can be visualized to show the relationality of the cultural objects they refer to, and thus, the “lifestyle of space” that Bourdieu maps in distinction