Sentiment and semantic analysis of Low-frequency NEED-passives on Twitter.

Avikshit Banerjee

Email: avikshit.banerjee@warwick.ac.uk (Avikshit Banerjee)

ABSTRACT

This study examines the low-frequency linguistic features of the English language through the verb NEED and its passive constructions. While previous sociolinguistic analyses have primarily focused on the regional syntactic variations of the three types of passives complements, namely, the past-participle complement (NEED+ED), the gerund participle (NEED+ING) and the infinitival (NEED+TO) complements, this examination sheds light on their semantic components and uses sentiment analysis to examine the relationship between these passive constructions and emotional affect in a Twitter corpus. Using two complementary sentiment analytic frameworks, we demonstrate that emotional valence and sentiments are significant regional dialect features which affect the production of two infrequent NEED-passive constructions, the (NEED+ED) and the (NEED+ING) complements and effectively study the factors affecting the decision processes of users for choosing between these constructions in a tweet. More generally, this project demonstrates the utility of sentiment analysis for examining low-frequency linguistic features and intends to facilitate future studies for other similar low-frequency verb passives.

Keywords: sentiment analysis, emotional affect, natural language processing, pretrained language model, transformer, alternate embedded passive, concealed passive, participle, gerund, regional dialects

INTRODUCTION

The English language corpora have majorly been studied through the lens of written and documented medieval literature and philosophy, followed by print text with the turn of the century and, most recently, through social media platforms. Several usage and linguistic peculiarities have manifested since, and these peculiarities are often characteristic of geographical locations and have become the subject of many dialectological studies of the English language (Edelstein, 2014, Murray, Frazer and Simon, 1996), especially concerning low-frequency, non-obligatory linguistic features (Strelluf 2022, Strelluf 2020).

One infrequent linguistic feature in standard English is the NEED-passive complement construction. The verb NEED can be used passively in the form of an infinitival where it is followed by a 'to' or a 'to be' (NEED+TO) known as the 'Standard Embedded Passive' (Edelstein 2014), a past participle (NEED+ED), called the 'Alternative Embedded Passive (AEP)' (Edelstein 2014), and a present participle or a gerund (NEED+ING) also called the 'Concealed Passive' (2002, cited in Edelstein, 2014, p. 244). Such is illustrated in (1) – (3), which are taken from tweets used in this study.

1. Still have 1 available that I need gone -(Columbus)

2. Do you have a specific question about technology that you need answered to find online forums and discussion board? -(Atlanta)

3. I need a proper tube bender and welder torch gas system. Need this to kick start rebuilding what I need to be done -(Singapore)

Past research by Strelluf (2022) and Strelluf (2020) on infrequent NEED-passives has tracked the variance in production for each passive form across dialectological regions and studied the syntactic constraints, hence adding more complexity to this previously neglected space. In his study, Strelluf (2022) takes advantage of the enormous speech-like text corpora on Twitter for analysing low-frequency features and establishes three distinct usages for the three NEED-participle forms across the world,

effectively tracing historical migration routes. However, one problem that keeps surfacing is the unexplained endurance exhibited by such features, especially by NEED-past constructions, despite being rarely used in conversations. In this study, I attempt to address this linguistic paradox and subsequently identify usage constraints that predict when each of the three passive constructions will be used among individuals who use multiple constructions, by quantifying its emotional semantics. For doing so, I will be utilising a corpus of tweets scraped from Twitter containing NEED-passive constructions from the United States, the United Kingdom, and other parts of the world and perform sentiment analysis by using the latest state-of-the-art Natural Language Processing (NLP) techniques, coupled with a conventional approach of quantifying various semantic parameters. The findings of this examination reveal that sentiments play a significant role in selectively licensing specific NEED-passive constructions and act as a reliable marker for dialectological boundaries. This study further intends to facilitate future investigation of the semantics of similar low-frequency verb passives like want, like, and deserve and subsequently offer potential insights for analysing and generating speech-like language.

BACKGROUND

This section provides an overview of previous research in the dialectological exploration of NEED-passives, including regional distribution and syntactic analysis of the three passive constructions. It introduces unresolved issues and positions sentiment analysis as a potential solution to these issues.

NEED - passives Historically, the most studied linguistic features have always been the frequent and obligatory ones, which were more of a compulsion than a choice due to their higher availability in the literature. Labov (2006:32) established that "the most valuable items are those that are high in frequency, have a specific immunity from conscious suppression, are integral units of larger structures, and may be easily quantified on a linear scale." Most of the literature has been restricted to documented or written corpora which don't include speech-like text, which is often more linguistically rich and involves infrequent linguistic idiosyncrasies.

Quantitative Linguistics has consistently argued that infrequent linguistic features constitute noise and hence are not reliable for dialectological and sociolinguistic study (Manning & Schutze 1999:199, Carver 1987:17). However, with the advent of social media platforms, opportunities for studying infrequent linguistic features opened up and subsequent works by (Strelluf 2020, Strelluf 2022), showed that including such features alongside the most frequent ones, result in a more enduring and coherent geographical clustering of dialectology. In his study, Strelluf (2020) identified the usage of the NEED-past participle to be a region-dependent linguistic feature and used Twitter to analyse this feature and established that the diaspora of the NEED-past form of usage had been mainly due to early immigrations by Ulster Irish and Scottish immigrants between 16^{th} to 18^{th} century. Despite being rarely used in conversations, the fact that NEED+ED continues to be a productive and enduring linguistic feature is still a mystery to which there seems to be no grounding. Strelluf (2020) points out that although the usage of the NEED+ED passive is temporally enduring, it is invisible psychologically, and most users do not acknowledge its presence unless explicitly pointed out in their speech. This provides further evidence for NEED+ED to be an unstigmatised feature of the English language, which operates somewhat "below the surface"(Strelluf, 2020).

Such a feature begs whether other elements of language are at play in maintaining such infrequent non-obligatory features. It hints at potential rules which may afford the user to use NEED+ED constructions or even allow the user to use NEED+ED and NEED+ING constructions interchangeably. Strelluf (2022), in his study, inferred that NEED+ED and NEED+ING passives might share the same derivational structure, which may differ from that of NEED+TO passives. His research found that both NEED+ING and NEED+ED passives mirrored each other in terms of regional variability. He postulated that having a similar derivational structure afforded the users the option to use either of the passives interchangeably. If true, this may account for the endurance of the NEED-past participle based on the adoption of the NEED+ING participle, which is more popular. However, as noted by Strelluf (2022), this observation may not account for the productive existence of NEED+ED, as there have been contradicting studies where it has been shown that although the AEP construction is sociolinguistically invisible, the users of the AEP, when presented with the NEED+ING

passive construction, rejected them as being ungrammatical (Murray, Frazer and Simon 1996:266) and disregarded the NEED+TO passive as being too formal. Another study (Doyle 2014) showed that the geographical diaspora of NEED+ED and NEED+ING were complementary in the United States.

This feeds the uncertainty concerning the rationale behind the NEED-past participle's continued endurance, despite no cognizance of its speakers. Given the diverse multidialectal space of the US Midlands in the early 18th Century, it wouldn't have been possible for children to cognitively process its relevance at that time as rarely ever occured (Strelluf, 2022). In this study, to further explore the uncertainty surrounding the NEED-past lexeme, I employ the latest advances in the Natural Language Processing domain along with a more conventional approach to perform sentiment analyses of the Twitter corpora and analyse the emotions involved with using low-frequency passive complements.

In Behavioural Psychology, emotional affect is well-known to have various effects on the cognitive latency of information retention. One such widespread behavioural phenomenon is the 'negativity bias'(Kahneman D, Tversky 1991, Alberini CM, 2010), which posits that emotionally negative information is more likely to leave a concrete and lasting subconscious impression than an equally positive counterpart, also famously advocated by the 'Prospect Theory' (Kahneman D, Tversky 1979) wherein "losses loom larger than gains". It might be possible that the emotional aspect of low-frequency linguistic features may be low in valence and high in arousal, which allows it to have a lasting subconscious impression which explains its endurance, observed over the centuries. In a study conducted by (Pratto & John, 1991), the participants seemed to remember the ink colours with more latency when it concerned words with low valence and high arousal, indicating that negative stimuli demanded more attention and focus (T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T., 1998). In these social experiments, the distinguishing feature is that the participants are provided with a visual trigger (the colours) for associating with the negative emotion. This allowed the participants to recall the stimuli upon command consciously. However, for NEED+ED passive speech, the negative emotional affect is unlicensed without any trigger, probably making them "psychologically invisible" (Strelluf 2020) yet subconsciously enduring.

Sentiment Analysis Sentiment Analysis is the phenomenon of computationally modelling text documents and assigning sentiments to them to categorise human emotions. It also has been a pivotal part of natural language research for almost a decade. With the advent of social media and the availability of vast corpora of speech-like text, sentiment analysis has become the go-to tool for analysing and predicting crowd sentiments. Most sentiment analysis applications have revolved around mining opinions related to consumer goods and products, movie reviews and social media speech. This approach has been highly successful in predicting opinions about political candidates, seasonal mood variation, and depression in medical patients and has been validated against the ground-truth ratings by the language producers themselves (Nguyen, Phung, Adams, Tran & Venkatesh 2010; Golder & Macy 2011).

METHODOLOGY

Twitter has historically proven to be an extremely resourceful source for conducting quantitative analyses of low-frequency linguistic features (Strelluf 2022, Strelluf 2020, Strelluf 2019). This is mainly because of the enormous volumes of speech-like text posted daily on its platform, which allows for even a tiny proportion of infrequent linguistic features to culminate into a significant number (Strelluf 2020, Eisenstein 2015, Eisenstein 2017).

Corpus Construction The dataset used in this study has been inherited from the Twitter corpus compiled by Strelluf (2022) for his research concerning the syntactic variation of NEED-passives (Strelluf 2022). The corpus comprised 55,261 tweets categorised under regions ('UK' for United Kingdom, 'US' for United States and 'world' for other international cities), cities the tweets were scraped from, tense (active and passive), the tweet-handles of the authors and finally the part-of-speech tags for each tweet ('ED' for NEED+ED, 'ING' for NEED+ING and 'TO' for NEED+TO constructions). However, it should be noted that no social information was collected about the authors of the tweets. Some examples of tweets are listed from (4) – (6).

4. "Still have 1 available that I NEED gone" -('ED', Columbus)

5. "Hey @mobibikes you NEED rebalancing immedi-

ately at second beach. Both stations are full with people waiting" -('ING', Vancouver)

6. "I NEED to be dragged screaming away from this tweet" -('TO', Aberdeen)

For this study, only the 'passive' tweets were filtered. Regular expressions were used to eliminate tweet-handles mentioned within the tweets (as they did not reflect the author's tweet handle), numbers and words exceeding twelve characters. The tweets were then tokenised and lemmatised using R packages 'tidytext' (Silge & Robinson 2016) and 'textstem' (Rinker 2018). Unnecessary stop-words were removed using 'tidytext"s built-in 1,149-word list of stop-words along with a custom list comprising the words 'need', 'needing', 'needs' and 'needed'. This ultimately resulted in a corpus containing 241,425 tokens.

Assigning sentiment ratings Two unique and complimentary sentiment analytic approaches were employed to analyse the tweets. The first method uses a specialised pre-trained language model (more under Appendix) known as the 'cardiffnlp/twitter-roberta-base-sentiment-latest', a state-of-the-art tweet sentiment classification model (more details under Appendix). The sentiment classification was performed with the help of their API (Application programming interface) known as the 'Inference API'. The sentiments assigned by this model were either 'Positive', 'Negative' or 'Neutral', and the scripts for performing the classification on each tweet were written in Python. Some examples of classified tweets are as follows:

1. "I NEED to be dragged screaming away from this tweet" -('NEGATIVE', Aberdeen)

2. "@kristynnz all good cryptic tweets NEED to be spoken about lol" -('POSITIVE', Auckland)

3. "@Girlw0nder Sort out what NEEDS to be sorted x" -('NEUTRAL', Liverpool)

As a prudent check for the sentiment classifier, I use another more conventional approach to perform sentiment classification, which involves assigning token ratings according to Osgood's semantic differential (Osgood, May & Miron 1975), which are based on valence (happy versus sad), arousal (engaged versus detached), and dominance (active versus passive), thus allowing our analysis to extend beyond the purview of only valence.

Osgood's semantic differential approach works by assigning each word a specific rating for valence, arousal, and dominance. These are then aggregated to calculate a tweet's net rating. One of the main drawbacks of this approach is that it is context-independent and polarity insensitive. This is especially problematic in dealing with negation words in a text (e.g., 'not')

For assigning these ratings, scripts were written in R for valence, arousal and dominance from Warriner, Kuperman, and Brysbaert's (2013) dictionary, which has ratings for almost 14,000 English words. However, a few of the tokens were not rated, mainly because not all tokens are present in Warriner, Kuperman, and Brysbaert's (2013) dictionary, thus resulting in many missing values, accounting for almost 7% of the total number of tokens. However, removing 7% of the sampled data may not be wise. Hence, I decided to impute the values using the MICE (Multivariate imputation by chained equations) package in R (Azur MJ, Stuart EA, Frangakis C, Leaf PJ, 2011) and using the 'predictive mean matching'(pmm) method for the missing values. For the ease of quantitative analysis, I converted the factor levels of Positive', 'Negative' and 'Neutral' to 1, -1 and 0, respectively.

In essence, we now have two unique sentiment classification methods. The 'RoBERTa'-based classifier resulted in polarity-dependent ratings, whereas Osgood's Differential ratings were polarity-independent such that each token was rated independently, outside the context of their corresponding tweets. For both the sentiment classification approaches, I performed an ANOVA(Analysis of Variance) using the 'afex' package in R between the participle forms and the norm ratings to check for any significant relationship. I also performed a multinomial logistic regression using the 'multinom()' function of the 'nnet' package in R between the norm ratings and the participle forms to investigate their dependence. This process was repeated for the participle forms and the sentiments assigned by the 'RoBERTa'-based sentiment classifier to analyse the relationship between the sentiments and the variance in participle usage.

I hypothesise that the tweets' latent sentiments will significantly affect low-frequency participle usage behaviour. The fact that the NEED+ED passive has dialectologically enduring may be due to its potential valence and sentiments. Given that lower valence (negative affect) or negative sentiments appear to increase cognitive latency of information and experience, I hypothesise that the NEED+ED and NEED+ING participle usage will be sig-

nificantly related to sentiments and valence and that the NEED+ED participle usage will be related to lower valence, higher arousal and concreteness ratings (Pratto & John, 1991) compared to the other participles.

RESULTS

We first examine the extent of variation in forms determined by the norm ratings by performing an ANOVA(Analysis of Variance) for each of the norm ratings. In **Figure** 1, we can see the variation of the different norm ratings concerning each form variable. In the ANOVA, all the norm ratings are highly significant, with p<.001 with F-scores = 50.62, 33.02, 29.33, 14.31, 24.41 for the norms Concreteness, Age of Acquisition (AOA), Valence, Arousal & Dominance respectively.

From the graphs in **Figure** 1, Concreteness seems to be decreasing significantly across the participles, being highest for NEED+ED and lowest for NEED+TO, the difference between NEED+ED and the rest being highly significant (p<.0001). Regarding Valence, the NEED+ED participle has the lowest valence, and NEED+ING has the highest in the entire dataset. However, it is essential to note that NEED+ED is significantly lower than NEED+ING (p<.0001), but the difference between NEED+ED and NEED+TO is barely significant (p = .02).

In terms of 'Age of Acquisition'(AoA) ratings, interestingly, NEED+ED has the lowest ratings in the entire dataset, which is significantly different from that of NEED+ING (p<.0001). NEED+TO seems to have the highest AoA ratings overall, which seems plausible given that NEED+TO is prescriptively the "correct" form and possibly the most "formal" (Murray, Frazer and Simon 1996:266).

For the arousal norm ratings, NEED+ING is substantially lower than NEED+ED and NEED+TO (p<.0001), similar to the valence norm ratings. In contrast, the difference between NEED+ED and NEED+TO isn't significant at all.

However, it seemed that both the Valence and Dominance rating variation across the participles were strikingly similar (**Figure** 1), which led me to analyse their relationship further. On performing a correlational study of the two norms, I found that they are strongly correlated with a Pearson Correlation Coefficient (r) = 0.7 and p<.0001. This was quite interesting as communication with a negative affect or lower valence, in general, is broadly associated with complaints, which can be assumed to have a reduced agency and hence lower dominance. In fact, studies have suggested social dominance as a significant factor in the complaint behaviour of people (Dina Abdel Salam El-Dakhs, Mervat M. Ahmed, 2021).

This eventually led to analysing the English vocabulary contained within Warriner, Kuperman, and Brysbaert's (2013) dictionary. Surprisingly, a very similar Pearson Coefficient of 0.71and p<.0001 emerged in the English corpora (**Figure** 2), which is fascinating as it essentially renders the 'Dominance' rating redundant. However, it would be interesting for future studies to investigate whether the same relationship holds for other languages or not, which, if it doesn't, might reveal something unique about that language itself. Unfortunately, due to its redundancy, I shall not be using the norm 'Dominance' further in the study.

In this study, however, we shall focus more on the 'Valence' and 'Concreteness' ratings for the NEED-passive participles to distil the specific emotional affect associated with the low-frequency linguistic features, if any, and possibly determine the factors and rules that govern the perceived noise in choosing participles in speech.

As speculated earlier, **Figure** 3 clearly shows that NEED+ED has the lowest valence ratings out of the three participle usages, supporting our primary argument. Least mean squares tests were performed for the three participle forms, which show highly significant differences between the means of NEED+ED and NEED+ING forms (p = 0.001 for 'UK', p<.0001 for 'The US' & p = .02 for 'world'). This observation provides significant momentum to this study's primary objective by revealing that NEED+ED constructions are indeed characterised by considerably low valence in general, potentially playing an essential role in the endurance of such low-frequency linguistic features.

We repeat the same analysis for all the other norm ratings for the three geographies (**Figure** 4) and observe that NEED+ED has the highest concreteness across all three geographies. This agrees with the previously discussed phenomenon that negative information appears more valid and immediate and tends to affect us more subconsciously. The usage of the NEED+TO participle is associated with the lowest 'Concreteness' rating, which might be due to the participle use being associated with more general and abstract topics. The difference between NEED+ED and

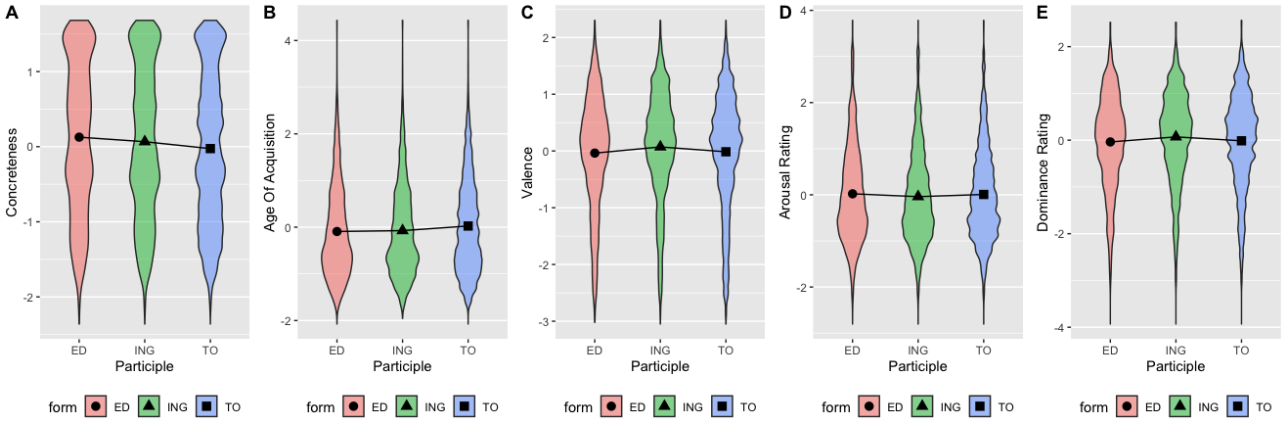

Figure 1. Shows the variation of the norm ratings concerning each form variable. The means are represented at the centre, along with the general distribution of the data around its mean.

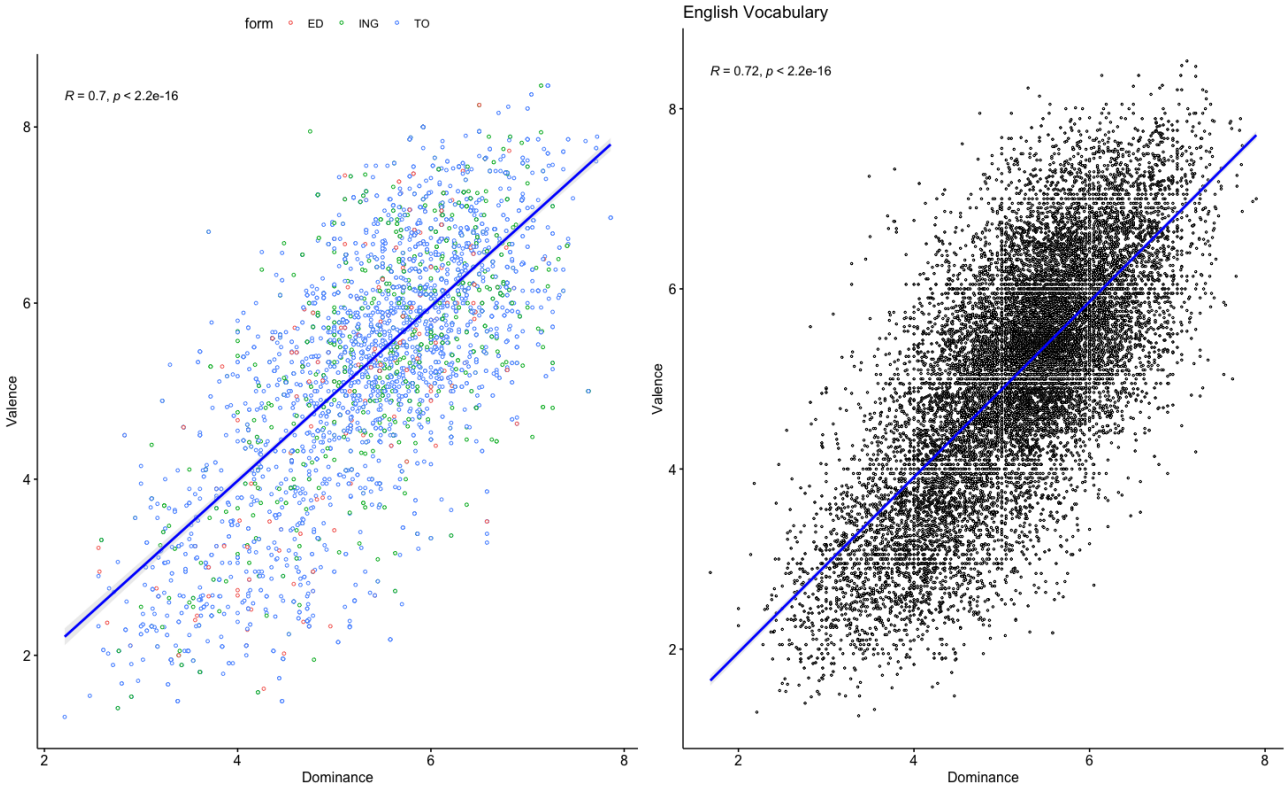

Figure 2. Correlational analysis between 'Valence' and 'Dominance'. (Left) The correlation between the two norms in our study sample. (Right) The correlational analysis between the two norms was performed on the entire English vocabulary.

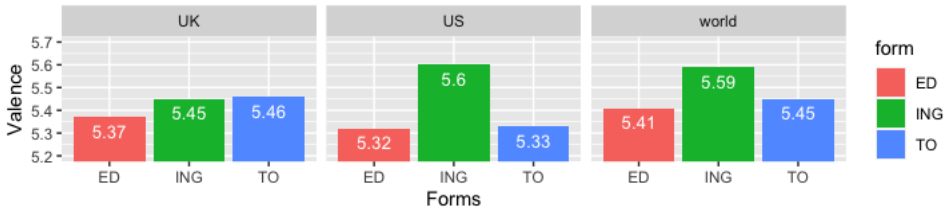

Figure 3. Mean Valence ratings aggregated across participles for three different geographies.

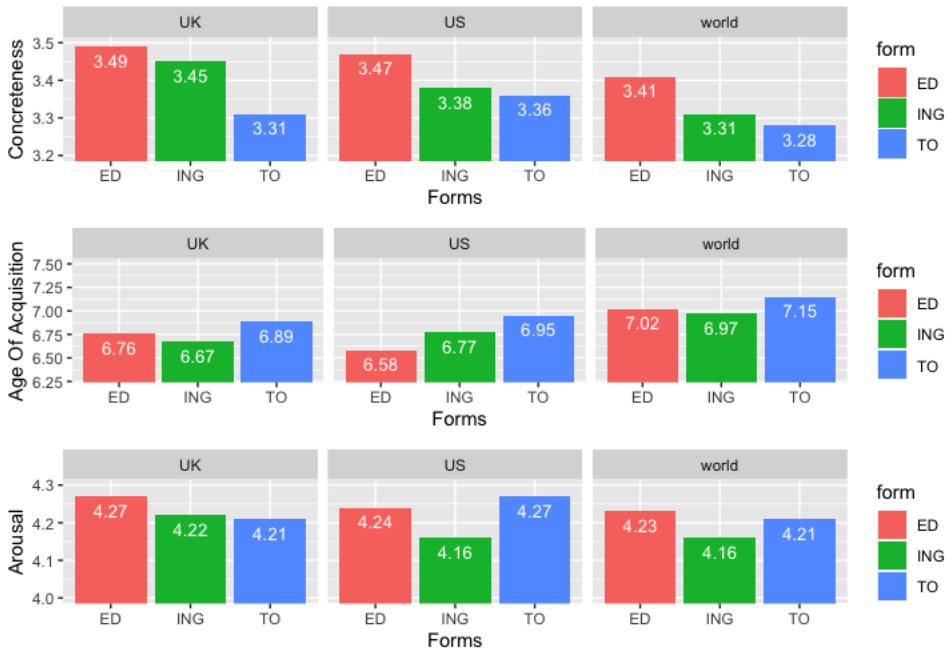

Figure 4. Mean Norm ratings across all participles for all three geographies.

NEED+ING is significant (p<.0001 in the UK, p= 0.0139 in the US & p = 0.023 for the rest of the world), and so is the difference between NEED+ED and NEED+TO for all three dialectological geographies (p<.0001 in the UK, p<.0001 in the US and p = .022 for the rest of the world). However, such isn't the case for NEED+ING and NEED+TO, where the difference isn't as pronounced in general.

Interestingly, for the 'Age of Acquisition' (AoA), the NEED+ED construction, despite its low frequency and rare use, had a significantly lower rating than the frequent NEED+TO construction (p<.0001 in general), especially in the US, which goes to answer a question earlier raised by Strelluf (2022), where he questions the endurance of such linguistic features, given that children in the early 1800s in the US Midlands, were probably not exposed to NEED-passives as a majority. From what we observed, it was very likely that many of the children, especially in the US, were exposed to NEED-passives, especially NEED+ED, shown by the significantly low AoA. This may be because children are more likely to be scolded and punished by parents than adults and hence are exposed to much lower valence words, characteristic of NEED+ED constructions. However, it is not significantly different from the AoA ratings for the NEED+ING construction, which, too, is substantially lower than the AoA rating of the NEED+TO construction (p<.0001) across all three regions. This supports Strelluf's (2022) claim that both NEED+ED and NEED+ING constructions may be derived from a similar structure and used interchangeably.

Not surprisingly, the mean arousal rating for the NEED+ED participle usage was significantly higher than that of NEED+ING (p<.0001 in general, p = .0002 in the UK, p<.0001in the US). However, under the 'world' region, NEED+ED is not significantly different in arousal ratings from the other two forms of usage. NEED+ING is markedly lower in arousal ratings than NEED+ED and NEED+TO (p<.0001), whereas NEED+ED and NEED+TO are not significantly different, as was expected due to their similar valence dynamics. Another interesting observation is that the similarity between the three forms of NEED-passives is geography-dependent in terms of valence and arousal. From **Figure** 4, in the US and the rest of the world, the NEED+ED passive and the NEED+TO passive differ significantly from the NEED+ING passive (p<.0001) but not from each other.

We now focus on the sentiment ratings generated by the state-of-the-art 'RoBERTa'-transformer (**Figure** 5). Once again, the sentiment ratings are consistently more negative for the NEED+ED passive constructions than the NEED+ING passives, both in general and within each dialect area. Also, as seen before, NEED+ING passives are observed to have the least negative sentiment involved with their constructions. They are significantly more posi-

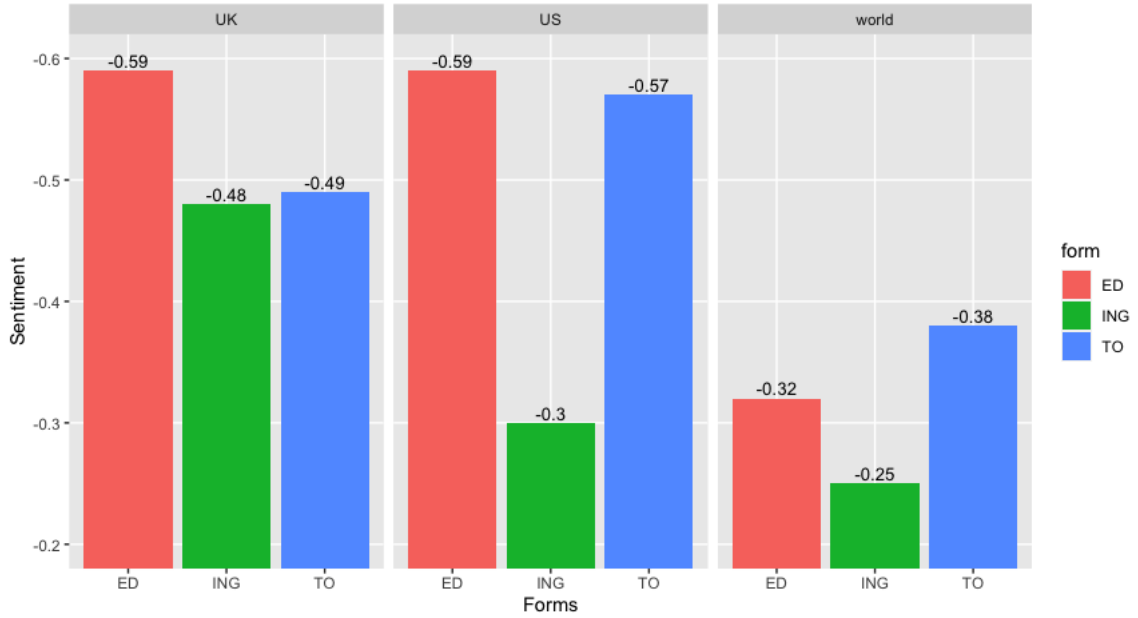

Figure 5. 'RoBERTa' Sentiment average ratings for all three geographies across all the participles.

tive than the NEED+ED passive constructions (p<.0001 in general, p<.0001 in the UK & p<.0001 in the US). However, as seen previously with the valence ratings in **Figure** 3, the mean sentiment difference is insignificant between NEED+ED and NEED+TO passive constructions for the dialect regions of the 'US' and the 'world'. In fact, in the UK, the ratings of NEED+ING and NEED+TO are not significantly different for either the valence ratings or the sentiment ratings (p = 0.70 for sentiments and p = 0.69 for valence ratings in the UK). It should be noted that the geographical dependency exhibited by the inter participle relationships are replicated in the sentiment analysis as well, with NEED+ED and NEED+TO lexemes being similar to each other and significantly different (p<.0001) from the NEED+ING lexeme in the US and the regions under 'world'. However, in the 'UK', the NEED+ING lexeme is similar to NEED+TO, and both are significantly different (p<.0001) from the NEED+ED passive.

From the above observations, it is clear that the usage of the NEED+ED passive construction is used with a more dispirited language, both in terms of having a lower valence and a more negative sentiment rating when compared to the NEED+ING passive, which on the other hand, is significantly associated with higher valence ratings and less negative user sentiments. The further complementary observation is that, in agreement with the valence and the sentiment ratings, the NEED+ED passive construction is significantly more concrete and tangible than the other forms, which, as discussed earlier, furthers the point that negative affect in text or information is perceived to be more valid and immediate as compared to positive information (Hilbig B., 2011). This pattern generally holds within each dialect area and sheds further light on low-frequency linguistic features' complex semantic nature. Interestingly, the NEED+TO passives and the NEED+ED passives do not differ significantly in the dialect regions besides the UK and the US, also referred to in our dataset as 'world', in terms of sentiment, valence, age of acquisition as well as arousal. This consistency in the relationship between the two NEED-passives in this region may imply that they might have a semantic similarity, despite having no apparent syntactic structural similarities (Strelluf 2022). It may be the case that in the post-colonial British colonies, where the usage of the NEED+ED passives is negligible, the NEED+TO passives are used as a semantic substitute.

In further analysis, I used a multinomial logistic regression model using the 'glm()' function to analyse the interaction of the various norm vectors and the sentiment ratings with the form variable. The results showed that, for every unit rise in the valence rating, the log-likelihood of a particular tweet having a NEED+ED passive construction rather than a NEED+ING one increased significantly by 0.042 (p<.0001). It also showed that with a single unit increase in the concreteness rating, the log-likelihood of the tweet being a NEED+ED passive rather than a NEED+ING one increased by 0.08, which, too, is highly

significant (p<.0001). This proves that the emotional valence, the nature of information and the dialect region are substantial contributors to the decision-making process of choosing between a NEED+ED and a NEED+ING passive construction.

I perform a bubble chart visualisation (see **Figure** 6) to visualise and trace the relationship between NEED+ED passive usage and the sentiments across the UK. In **Figure** 6(B), NEED+ED participle usage is mainly focused in the northern parts of the country, mostly in the cities of Glasgow, Edinburgh, and Belfast, followed by Aberdeen and Newcastle. This is also per the results published in Strelluf (2020). Interestingly, the 'RoBERTa' sentiment bubble map representation of the country agrees broadly with the usage proportions. We can see that the negative sentiments are the highest in the north (mostly in red, orange and yellow), with a sharp decline towards the south, especially after the English Newcastle, which also happens to serve as the southern border for NEED+ED participle usage in the UK (Strelluf 2020). However, Norwich seemed to be an exception as it also scores highly in negative sentiments. All the northern cities with a high proportion of NEED+ED usage also seem to be high on negative sentiments, hence adding power to our argument that sentiments significantly determine the use of low-frequency linguistic features.

However, it is also observed that Osgood's differential Valence ratings are not as deterministic as the 'RoBERTa'-based language model in reflecting the NEED+ED usage proportions. This might be due to its failure to account for polarity in tweets, as it mainly depends on aggregating the valence ratings for each tweet without accounting for context. However, we can still see that the negative valence ratings are higher in the northern parts of the country with an overall southward decline. Like the sentiment analysis, Belfast has the highest negative valence ratings overall. The valence bubble chart still retains the general negative character of the north compared to the south, which is also consistently reflected in the NEED+ED usage proportions.

We repeat the above visualisation with the cities in the 'US'. From **Figure** 7(A & B), it is observed that the 'RoBERTa'-based sentiment visualisation once again traces the NEED+ED usage footprint across the geography, with the sentiments being the most negative towards the eastern cities of Pittsburgh, and Columbus, followed by a steep decline towards the west with Los Angeles and San Francisco having the most positive sentiments. This traces the NEED+ED usage proportions, which are the highest in the eastern US midlands of Pittsburgh and Columbus, due to the early Ulster Irish and Scottish immigrants (see Strelluf, 2020), with a steep decline from Kansas City and westward, hence following the linguistic imprints of migrations in the early seventeenth century and still being the basis for distinct dialectological boundaries. Strelluf (2020), in his study, recorded that NEED+ED usage proportions do not diffuse eastward from Pittsburgh to Philadelphia, which is within 300 miles but do so towards Kansas City, roughly 850 miles to the west. Our visual representation of the sentiment ratings (**Figure** 7 B) also replicates this finding.

Similar visual analyses were also carried out using ING participle and the geographical variation in sentiments and mean valence ratings. As such, it has been outlined previously by Strelluf (2022) that NEED+ING is a distinct feature of British English, especially in England and Wales, and this can be seen in **Figure** 9(B). The interesting observation here is the sentiment bubble chart (**Figure** 9A) which also maps a similar spread for positive sentiments, the most positive sentiments being centred in England and a westward spill over to Wales, which agrees with our previous statistical finding of NEED+ING passive usage being significantly associated with higher Valence ratings and positive sentiments. It is also interesting to note that Newcastle acts as a bridge between northern and southern England due to its moderated sentiment rating values, as seen in **Figures** (6A and 9A). This also adds credence to the fact that Newcastle-upon-Tyne is the only English city that uses NEED+ED and NEED+ING structures on an equal footing (Strelluf 2022) and further bolsters the fact that sentiment ratings do indeed play a significant role in defining dialect boundaries and significantly mirrors the usage of NEED-passive complements.

In **Figure** 10, we analyse the same for the US cities. Here too, we can see the sentiment bubbles mirroring the diaspora of NEED+ING participle usage in the states, with the most positive sentiments being associated with the easternmost cities of Boston, New York, and Philadelphia, followed by the western cities of San Francisco, Los Angeles, and New York (**Figure** 10B). This geography is repeated with the NEED+ING passive usage proportions as seen in **Figure** 11C and Osgood's Valence ratings in **Figure** 11D, adding more credibility to our primary argument.

On performing a correlational analysis between the pro-

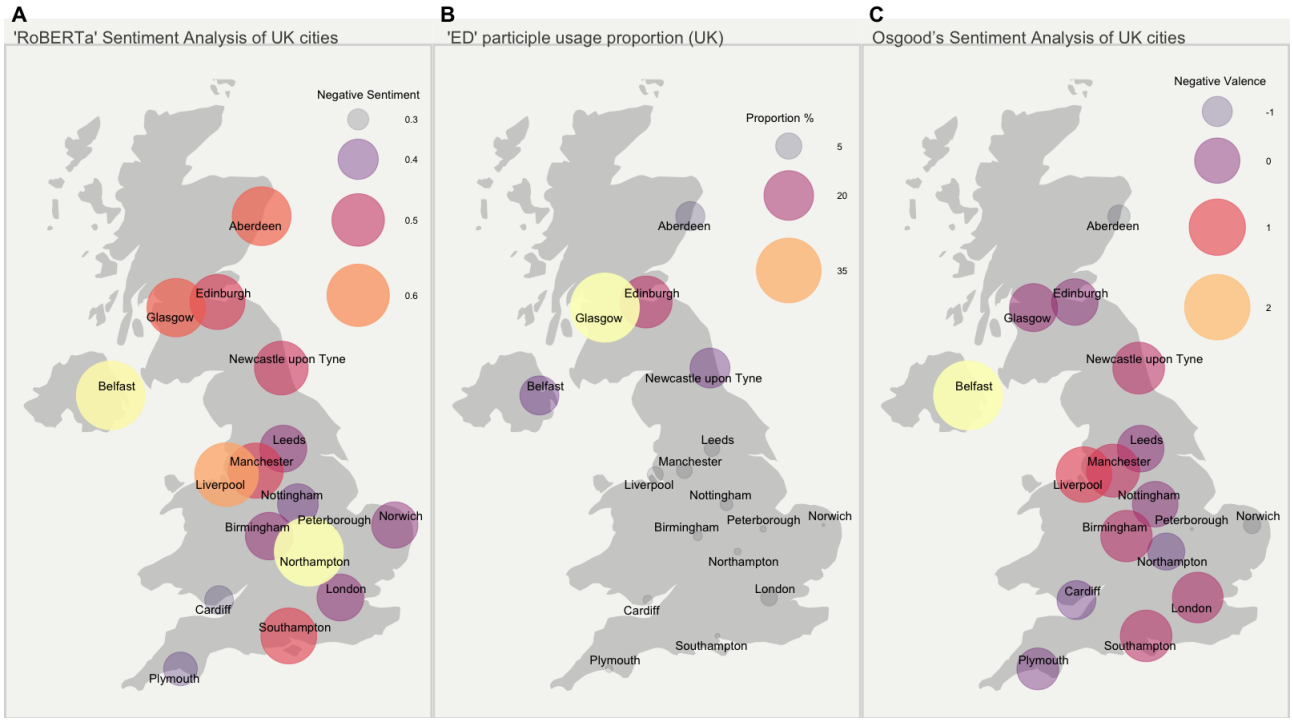

Figure 6. A bubble chart visualisation of the UK (A) showing the sentiment variation across the UK (B) representing the proportion of ED usage (C) showing the variation of Valence ratings of tweets across the country.

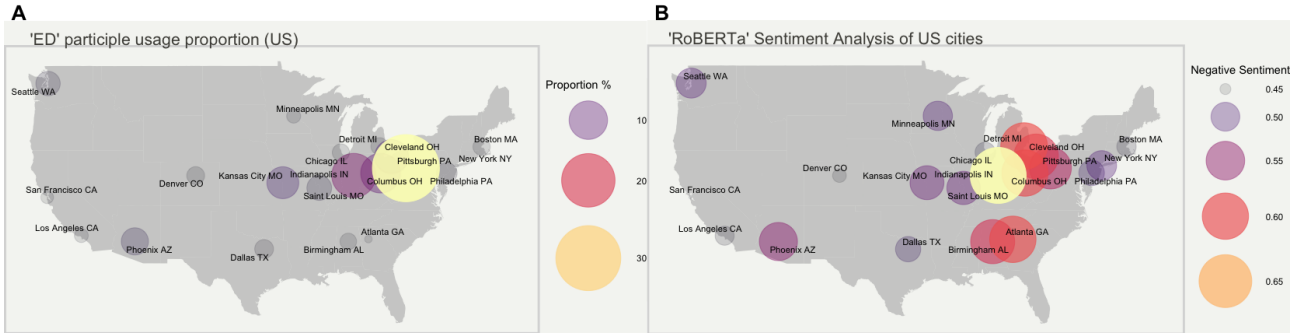

Figure 7. A bubble chart visualisation of the US (A & B) showing the comparison between geographies using NEED+ED passives along with the geographical variation in mean sentiments

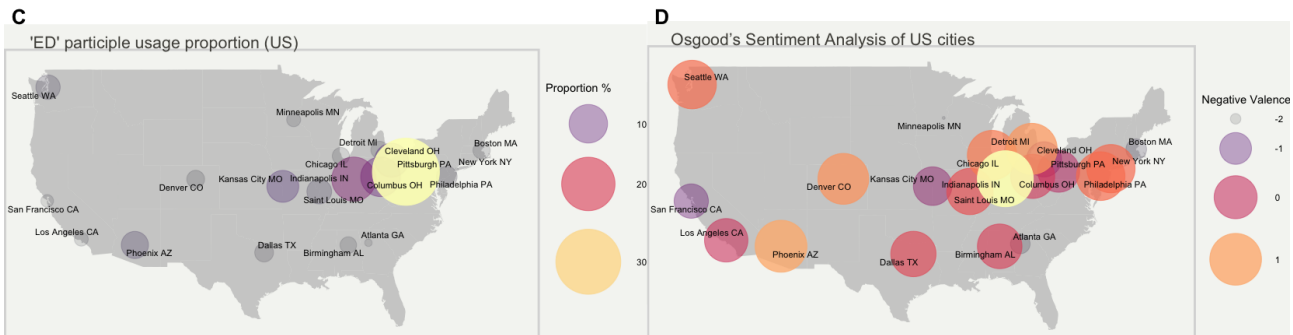

Figure 8. A bubble chart visualisation of the US (C&D) showing the variation of geographical variation of Valence along with the NEED+ED usage geographies with proportions.

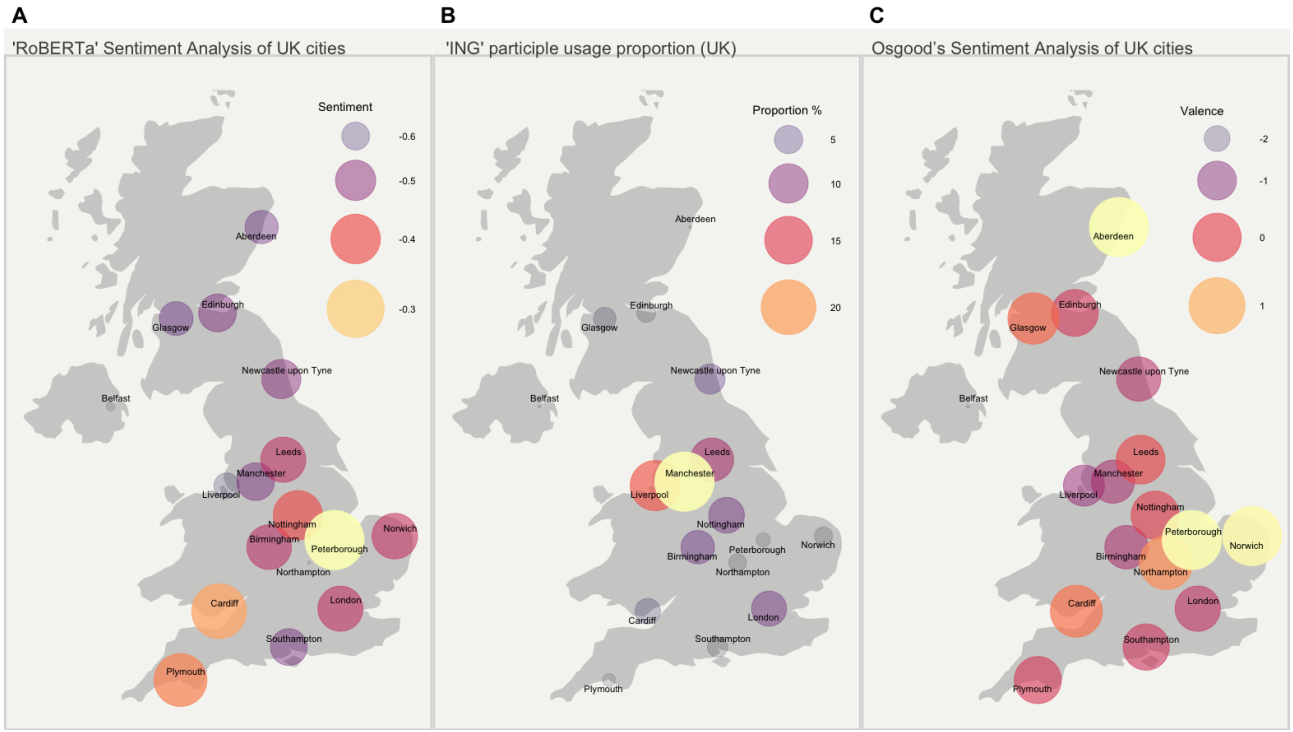

Figure 9. A bubble chart visualisation of the UK (A) showing the sentiment variation across the UK (B) representing the proportion of ING usage (C) showing the variation of Valence ratings of tweets across the country.

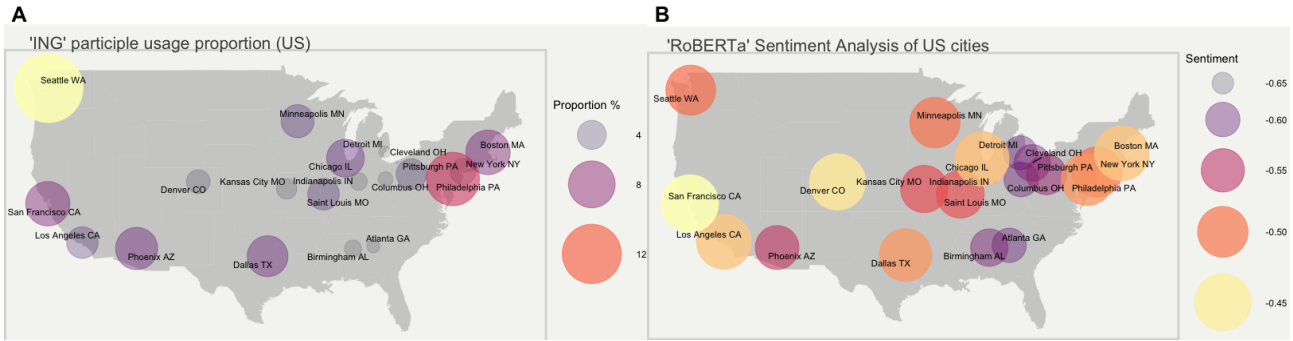

Figure 10. A bubble chart visualisation of the US (A & B) showing the comparison between geographies using NEED+ING passives along with the geographical variation in mean sentiments,

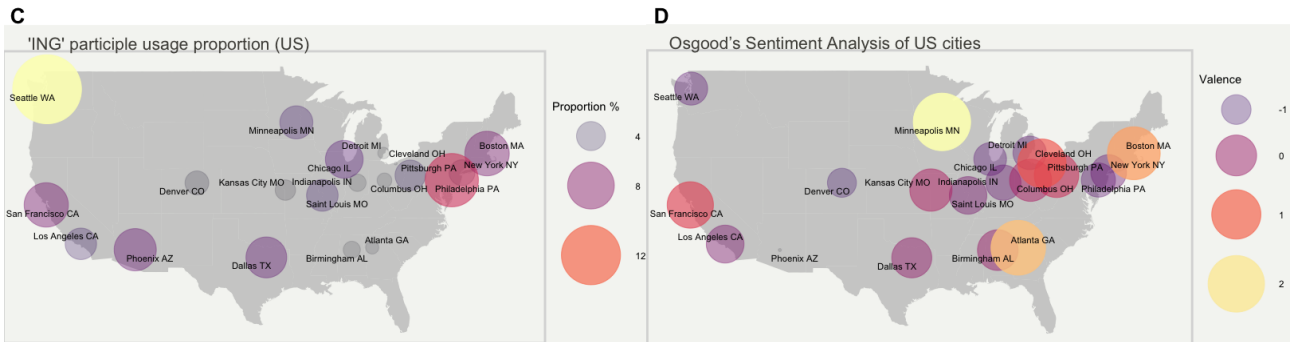

Figure 11. A bubble chart visualisation of the US (C&D) showing the variation of geographical variation of Valence along with the NEED+ING usage geographies with proportions.

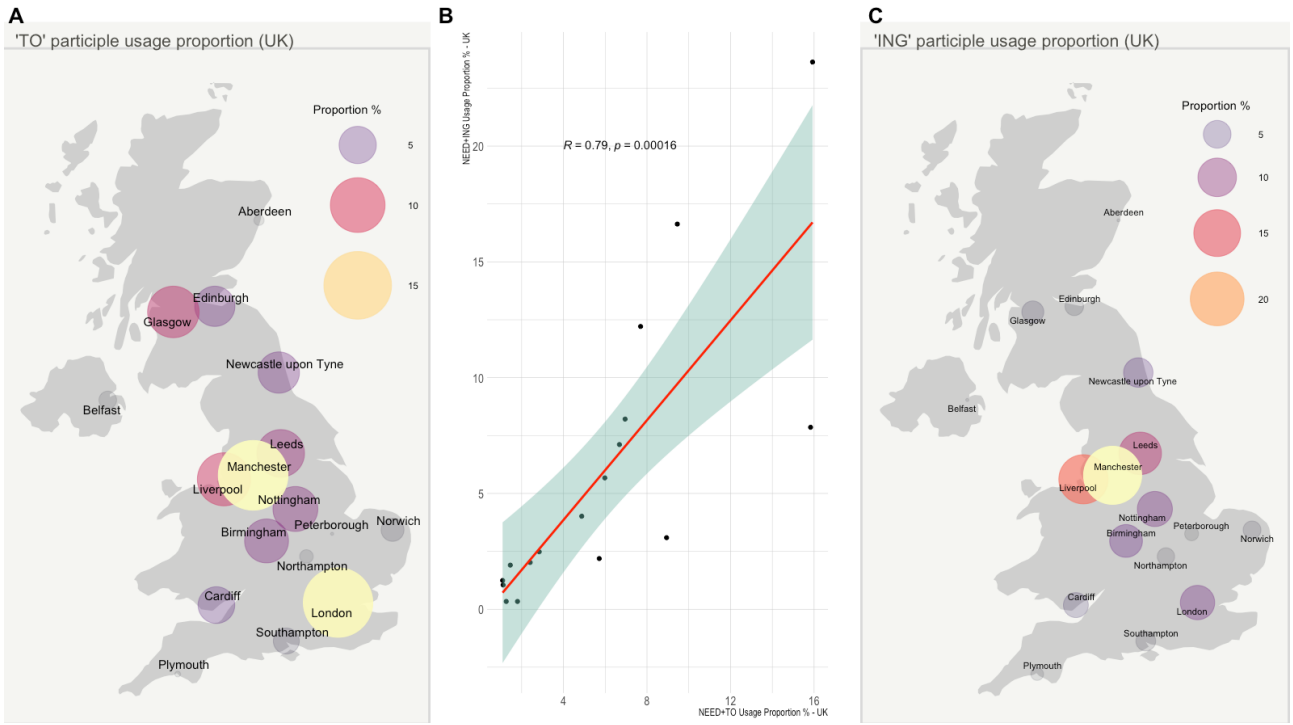

Figure 12. (A) Bubble plot showing the 'NEED+TO' participle usage distribution in the UK. (B) A plot showing the relationship between the usage patterns of the 'NEED+TO' participle and the 'ING' participle. (C) Bubble plot showing the 'NEED+ING' participle usage distribution in the UK.

portion distributions of the NEED+ING passive and the NEED+TO passive complements in the UK, we found that they are significantly correlated (Pearson Coefficient = 0.79, p<.001, **Figure** 12B). This is extremely interesting as it suggests a hidden dependency between the less frequent concealed passive and the more popular standard embedded passive. This can also be seen visually in **Figure** 12, that the NEED+ING usage proportion broadly mirrors that of the NEED+TO participle. Contrary to what Strelluf (2022) had suggested in his study, it may be the case that NEED+ING and NEED+TO passive complements share the same semantic structure and hence are used interchangeably, as can also be seen from the correlational and geographical analysis, and this, in turn, ensures the sustenance of a relatively low-frequency linguistic feature, i.e., the NEED+ING passive, independent of the NEED+ED passive, in the UK.

DISCUSSION AND CONCLUSION

This study established that the infrequent linguistic features of the NEED-passives are significantly dependent on the emotional affect of speech, and their interdependencies are predicated upon complex factors such as semantic variation and dialectological regions. It has revealed that NEED+ED constructions are fundamentally associated with lower valence, higher arousal, higher concreteness, and negative sentiments, as confirmed by two complementary sentiment analytic approaches. These outcomes demonstrate claims of the 'negative bias' attributed to the works of Kahneman D, Tversky 1979 and Alberini CM, 2010. Claims made by Strelluf (2020) and Strelluf (2022) concerning the efficacy of the NEED-past participle serving as enduring dialectological boundaries are further bolstered by the establishment of lexical sentiments and mean valence as significant regional features, which in turn affords the efficient production of NEED+ED lexemes as a productive and enduring feature. The observation that the sentiments of speech are the fundamental licensors of patterns in low-frequency linguistic behaviour has subsequently helped to answer long-standing questions surrounding the unexplained preservation of infrequent linguistic features despite being psychologically invisible to the masses and rarely used in regular conversations (Strelluf 2020, Strelluf 2022).

The negative emotional affect of NEED-past sentences makes for a case of lexical semantic negativity rather than grammatical negativity. This leads to the non-requirement

Figure 13. (Left) NEED+ED participle word cloud; (Middle) NEED+ING participle word cloud; (Right) NEED+TO participle word cloud. All the words have been sourced from the tweets used in this study.

of a polarity licensor in speech or text to express negative affect or emotion. I assume that this peculiarity affords the user to employ affect-licensing as a substitute for polarity-licensing for expressing a non-positive sentiment, i.e., it allows the user to engage negative affect without the explicit inclusion of a trigger word. This unique feature possibly affords it to go unnoticed by users and operate 'beneath the surface.' (Strelluf 2020) while still being an enduring regional grammatical feature. A general idea of the words associated with each NEED-passive complement can be observed from the word cloud associations in **Figure** 13, where the words coloured 'red' are the most frequent, followed by 'yellow' and the least frequent words coloured in 'grey'.

The remarkably significant correlation between the usage dynamics of the NEED+ING and the NEED+TO passives in the UK (see **Figure** 12) calls for considerable attention and further study, as it may represent hidden semantic similarities between the two passive constructions, which is contrary to what Strelluf (2022) had proposed earlier. From this finding, it can be assumed that the sustenance of the relatively infrequent gerund NEED-passive (NEED+ING) is hooked onto the usage dynamics of NEED+TO passives in the UK due to their significant semantic similarities concerning valence and sentiments. Contrary to Strelluf (2022)'s proposal, I speculate that the endurance of the NEED+ED and the NEED+ING constructions are rooted in underlying lexical semantics, which are independent of each other. This is further emphasised by their significant semantic differences in mean valence, sentiment, concreteness, and arousal, as investigated in this study. This relationship also explains the absence of NEED+ING construction in Scotland and Ireland, where the general underlying sentiment is significantly more negative than in England. This naturally favours the production of the more semantically suited NEED+ED passive construction and suppresses the subsequent production of the infrequent concealed passive. However, the NEED+TO passive continues to be a part of Northern UK and Ireland, despite its significantly positive semantic dynamics, as it is unable to be suppressed by virtue of its sheer frequency of production, albeit its usage is visibly reduced in the north as compared to the midlands and the south.

In this study, we also observe the differences in performance and efficacy of the two sentiment analytic approaches employed. The 'RoBERTa'-based pre-trained transformer proved to be more accurate in tracing the dialectological boundaries, while Osgood's differential ratings were not as accurate in showing the complex regional variance of valence ratings. From the 'Results' section, we know that NEED+ED usage proportions were highly significant in determining 'Negative' sentiments ($p<.0001$), which is visually evident as well in **Figures** 6(A&B) & 7(A&B). This might be due to the polarity-sensitive nature of the model and its novelty in classifying tweet sentiments by its state-of-the-art pre-trained architecture.

I speculate that similar semantic and sentiment analyses of other infrequent linguistic features will reveal similar complexities in their dynamics. At the least, in this study, I could demonstrate the utility of sentiment analysis in providing a possible solution to the long-standing questions surrounding the established durability of infrequent linguistic features and as a significant and robust regional

dialectological variable.

APPENDIX

The following sections provide a context for understanding the latest advancements in the NLP space for performing sentiment analysis and give a brief background for terminologies that will be used to discuss the sentiment analysis process going forward.

Language Models

Language models are fundamentally probability distributions over a vocabulary of tokens or words in a particular language. These dictate the conditional probability of a word/token occurring in a phrase or sentence, given a specific preceding word. Such models are generally used for text or speech generation, auto-completion, email grammar correction, and other text-based NLP applications. The model we use in this study is a sub-class of a language model known as a 'Transformer' (more under Appendix), a state-of-the-art NLP architecture. To be specific, we use the latest 'RoBERTa'-based Transformer architecture, which is designed explicitly for tweet sentiment classification.

Neural Networks

Neural networks are probably the most fundamental building block for almost every application employing Artificial Intelligence, and a Deep Learning framework is the Neural Network Architecture. Neural networks are a network of computational nodes designed to mimic the working of neurons in the human brain broadly. Since its discovery, the Artificial Neural Network (ANN) has become one of the most widely used architectures for Natural Language Processing.

Figure 14 represents a very simplistic schematic structure of an artificial neuron which, in essence, mimics the working of a natural biological neuron. The input-output process typically performs a weighted linear function on the input signals, which are then either used as is or are used for a sigmoid computation to compute probabilities or used as a threshold value beyond which the output signal may be passed on to other units (For details, please see Hinton, G. E., 1992). A network of such neurons is generally called 'Feed Forward Network' and is primarily used with word vectors for either predicting the next word in text sequence or classifying sentiments.

Convolutional Neural Network (CNN)

Convolutional Neural Networks or CNN (LeCun et al., 1998) are a modified ANN architecture specialised for pattern recognition in sequential data sets and image recognition in images. It helps to factor observable patterns in data into the established ANN architecture. CNNs have a hierarchical architecture and generally are composed of three essential layers: the convolution layer, the pooling layer, and the fully connected ANN layer (**Figure** 15).

Due to their hierarchical architecture and matrix representation of sentences, they have been generally preferred for performing sentiment classification. Studies have also shown gated CNNs to outperform many other architectures in terms of performance, especially in the field of NLP (Yin et al., 2017).

Recurrent Neural Networks (RNN)

A Recurrent Neural Network (RNN) (Elman, 1990) is a modified version of a feed-forward neural network such that a single node forms a connection between nodes in a temporal sequence. These are specifically suited for processing sequential data as data is fed into a node in the sequence. At each temporal time step, the output from the last element in the series is fed as input along with the next part in sequence to the same node. This temporal scope of an RNN provides it with the additional ability to possess short-term memory and can better deal with position invariance, which is not applicable for feed-forward networks (Mikolov, Tomas, et al., 2010) and hence makes it a robust architecture for capturing short term context windows in text sequences.

Figure 16 shows a simplistic schema for an RNN-based LM. The architecture comprises an input layer 'x', a hidden layer 'h' (context layer) and an output layer 'y'. The input word-vector representation x(t) forms the input to the context layer 'h' at a time 't' along with the context vector, context (t-1), from time t-1.

This architecture is well suited for short sequences as the performance drops for arrangements beyond 5-6 grams (Mikolov, Tomas, et al., 2010). However, this arrangement forms the fundamental building block of the language model that we employ in this study, for performing sentiment classification of Tweets.

Transformers

'Transformers' are a relatively new architecture based on previous CNN-RNN-based architectures, but without serial computation restraints and with long text sequence modelling capabilities. It employs the new state-of-the-

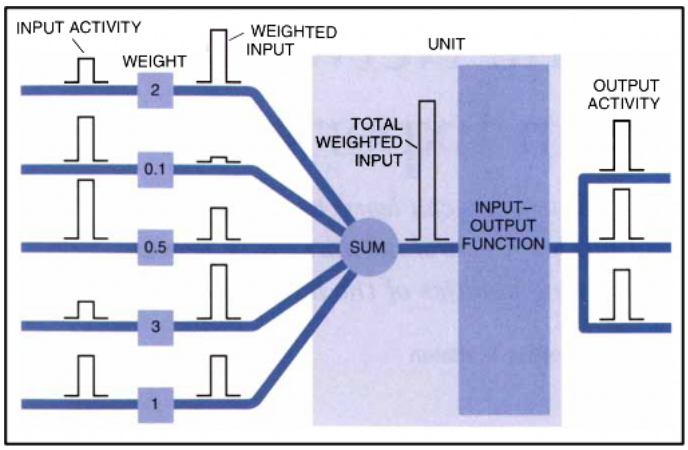

Figure 14. A schematic structure of an artificial neuron Source: Hinton, G. E. (1992). How Neural Networks Learn from Experience. Scientific American, 267(3), 144–151. http://www.jstor.org/stable/24939221

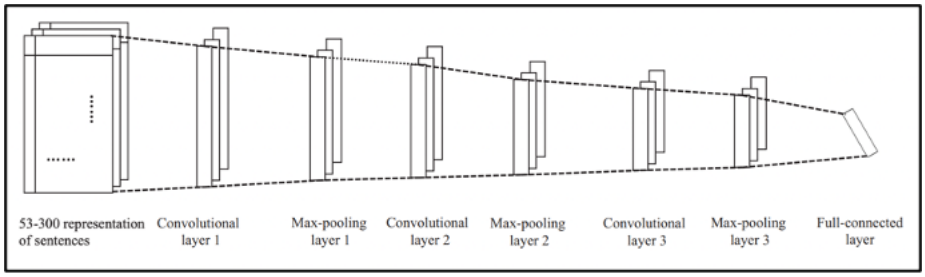

Figure 15. A simplified CNN architecture Source: (X. Ouyang et al., 2015, Figure 1)

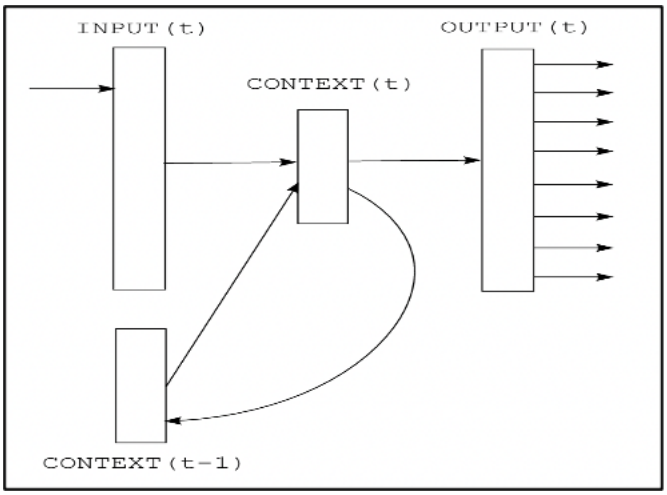

Figure 16. RNN-based Language Model, source: (Mikolov, Tomas, et al., 2010)

art‘ Attention Mechanism’(Vaswani et al. 2017, Section 3.2) function which gives the language model an unprecedented understanding of long-term contexts in text documents coupled with parallel computation capabilities.

The Transformer consists of two parts primarily, the encoder and the decoder. The encoder is an ANN-based process which essentially converts each text element in an input sequence into a corresponding vectorised form (S.T. Kokab et al., 2022). The idea is that the more similar the words are semantically, the closer their vectors or ‘embeddings’ will be in the vector space. On the other hand, the Decoder uses these embeddings for various NLP tasks such as Question Answering, speech translation and sentiment analyses (BERT is used for sentiment analyses, more on that later) as these embeddings capture the entire context of the input sequence. At a very high level, the novel

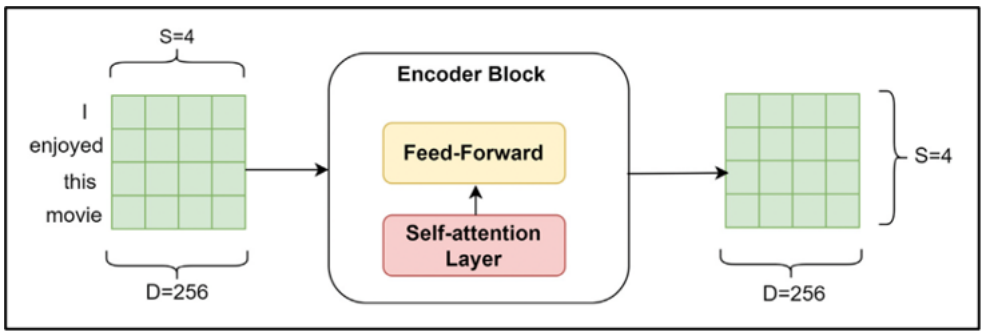

Figure 17. BERT architecture Source: (S.T. Kokabet al., 2022, Fig1 and Fig 2)

'attention mechanism' of transformers can be looked at as a sophisticated function which informs the model about all the past contextual words that need to be considered, while processing the present word. This very feature allows for extremely powerful semantic and sentiment analytic capabilities, especially in text documents where there might be complex polarity structures involved.

For our study, we use a BERT-based transformer language model (**Figure** 16) for performing sentiment analysis which has been specifically trained to analyse and classify sentiments of tweets. 'BERT' stands for 'Bidirectional Encoder Representations from Transformers.' It is a pre-trained language transformer developed by Google in 2018 (Devlin et al., 2019). Here 'Pre-trained' refers to the transfer learning process, which involves using machine learning to transfer knowledge gained in one novel application to another (S.T. Kokab et al., 2022).

From the architecture shown in the above picture, we can see that the Input sequence is treated as a matrix of vectors (each vector for each input element) due to parallelisation, which is then processed by the 'Self Attention Layer' (Vaswani et al. 2017, section 4) which produces a novel contextual heat map for the entire sequence which is then fed into a feed-forward neural network architecture which in turn creates a word embedding matrix for the whole of the sequence representative of the whole 'Bi-directional' semantic of the entire text sequence.

What makes BERT different from other transformers is its ability to learn and understand the linguistic context by the merit of its stacked encoder architecture, which then allows it to apply this sophisticated knowledge of language and context to a variety of applications including sentiment classification of texts in various languages.

The model that we utilize in this study for sentiment classification is called 'RoBERTa' which stands for Robustly Optimized BERT Pretraining Approach. It is essentially a modification of the existing BERT transformer architecture with a few functional changes in its design principles. To achieve higher performance with the existing BERT architecture, the model has been pre-trained for more extended periods, with bigger batches of data on longer sequences (Liu et al., 2019). It is pre-trained on 124M tweets scraped between January 2018 and December 2021 (Loureiro et al., 2022), and fine-tuned for performing sentiment classification as per the 'TweetEval' (Barbieri et al., 2020) benchmark with the most up-to-date trends in social media for identifying sentiments with extraordinary accuracy, which may not have been possible with more conventional approaches, given that the social media speech corpora and polarity interpretation change rapidly over time.

REFERENCES

European chapter of the Association for Computational Linguistics, pages 98–106.

SIGIR Workshop on Neural Information Retrieval.

Cognitive Science, 14(2):179–211.

change in social media. *PLoS One*, 9(11).

Handbook of dialectology. pages 368–383. Wiley.

Micro-syntactic variation in North American English. Oxford: Oxford.

(2016).

Alberini, C. M. (2010). Long-term Memories: The Good, the Bad, and the Ugly. *Cerebrum*, PMCID:3574792–3574792.

& David Silge, J. and Robinson (2016). tidytext: Text mining and analysis using tidy data principles in R. *The Open Journal*, 1(3).

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Barbieri, F., Camacho-Collados, J., Anke, L. E., and Neves, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Cacioppo, J. T., Cacioppo, S., and Gollan, J. K. (2014). The negativity bias: Conceptualization, quantification, and individual differences. *Behavioral and Brain Sciences*, 37(3):309–310.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018).

Doyle, G. (2014). Mapping dialectal variation by querying social media. *Shuly Wintner*.

Edelstein, E. (2014). *This syntax needs studied*.

Eisenstein, J. (2017). *Written dialect variation in online social media*.

Eisenstein, J., Connor, B. O., & Eric, N. A. S., and Xing, P. (2012).

El-Dakhs, D. A. S. and Ahmed, M. M. (2021). A variational pragmatic analysis of the speech act of complaint focusing on Alexandrian and Najdi Arabic. *Journal of Pragmatics*, 181:120–138.

Elman, J. L. (1990).

English, A. *American Speech*, 71(3):255–271.

Gentry, J. (2015).

Golder, S. A. and Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881.

Hilbig, B. (2011). Good things don't come easy (to mind): Explaining framing effects in judgments of truth. *Experimental Psychology*, 59(1):38–46.

Hinton, G. E. (1992). How Neural Networks Learn from Experience. *Scientific American*, 267(3):144–151.

Ito, T. A., Larsen, J. T., Smith, N. K., and Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology*, 75(4):887–900.

Kahneman, D. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, 47:278–278.

Kahneman, D. and Tversky, A. (2013). Prospect theory: An analysis of decision under risk. *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127.

Kokab, S. T., Asghar, S., and Naz, S. (2022).

Labov, W. Sharon ash & charles boberg. 2006. the atlas of.

Labov, W. (2006). *The social stratification of English*. New York City; Cambridge.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, . ., and V (2019).

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., and Camacho-Collados, J. (2022).

Manning, C. and Schutze, H. (1999).

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. *In Interspeech*, 2(3):1045–1048.

Murray, T. E., Timothy, C., & Beth Lee Frazer, and Simon (1996).

Nguyen, T., Phung, D., Adams, B., Tran, T., and Venkatesh, S. (2010). Classification and pattern discovery of mood in weblogs. In *Pacific-Asia Conference*

on Knowledge Discovery and Data Mining, pages 283–290. Springer.

Osgood, C., Egerton, W. H., May, Murray, S., and Miron (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press, Urbana, IL.

Ouyang, X., Zhou, P., Li, C. H., and Liu, L. (2015). technology; ubiquitous computing and communications; dependable, autonomic and secure computing. *IEEE international conference on computer and information*, pages 2359–2364.

Phonetics and Berlin.

Riezler, S. G. S., editor. *Proceedings of the 14th conference of the*.

Rinker, T. W. (2018). textstem: Tools for stemming and lemmatizing text version 0.1.4.

Strelluf, C. (2020). *Needs+PAST PARTICIPLE in US and UK regional Englishes on Twitter. World Englishes*, 39(1):119–134.

Strelluf, C. (2022). Regional Variation and Syntactic Derivation of Low-frequency need-passives on Twitter. *Journal of English Linguistics*, 50(1):39–71.

Sweden.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):1039–1061.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, . ., and I (2017).

Wen, Y., Zhang, W., Luo, R., and Wang, J.

AUTHOR BIOGRAPHY

Avikshit Banerjee u2147902