The influence of semantic features on lexical geographical variation

Abstract—In this paper, we investigate the influence of semantic concept features on lexical geographical variation. More specifically, we take an onomasiological approach to inquire into the effect of concept vagueness, salience, affect and semantic field. We use quantitative operationalizations of these features as predictors in a linear regression analysis. Our response variable is a composite variable which takes into account the number of variants per concept and the degree to which the concepts are scattered across geographical space in a heterogeneous way. Our model reveals that concepts show significantly more variation and that the lexical variants for these concepts are scattered across geographical space in a less homogeneous way. We also find differences between semantic fields.

Keywords—dialectometry, lexical variation, Dutch, quantitative lexicology

I. BACKGROUND

In dialectometry, language variation is often assumed to be governed by lectal or geographical factors (Nerbonne & Kleiweg, 2003; Séguy, 1971; Wieling, Nerbonne, & Baayen, 2011). However, a pilot study on the semantic field the human body showed that semantic features can influence lexical geographical variation as well (Author & author, 2010; Author & author, 2009). More specifically, the pilot study provided significant evidence for the influence of concept vagueness, salience and negative affect: vaguer concepts, such as LIES (‘groin’) often show more lexical geographical variation than less vague concepts, such as DUIM (‘thumb’); less salient concepts, such as SLUIK HAAR (‘straight hair’) show more lexical geographical variation than more salient concepts like HOOFD (‘head’) and negatively connoted concepts, such as KWIJL (‘drool’) show more lexical geographical variation than neutral concepts like JUKBEEN (‘cheekbone’).

II. RESEARCH QUESTIONS

In this paper, we expand on the pilot study in two ways. First, we expand the scope to other semantic fields than the human body. As a result we are able to show that, on the one hand, the influence of concept features on lexical geographical variation is relatively stable across different semantic fields. On the other hand, we can also determine the influence of the semantic field itself on dialectal variation. Methodologically, we also aim to take alternative operationalizations of the predictors that were used in the pilot study, viz. vagueness, salience and negative affect, into account. This allows us to compare which quantitative measures serve as good operationalizations of these predictors.

III. DATA & METHODS

For our analysis, we use the digitized database of the Dictionary of Limburgish dialects. As this onomasiological dictionary is divided into different semantic fields, we can determine whether the semantic field of a concept influences the degree to which the concept is prone to lexical geographical variation.

We use quantitative operationalizations of the semantic features, concept vagueness, salience and affect, as the predictors in a linear regression analysis. More specifically, we operationalize concept vagueness as the number of lexical types per concept that occur for other concepts as well. We calculate concept salience by taking into account the proportion of multi-word responses, the relative number of places without a response per concept in our data set and an operationalization of the prevalence per concept, using data that was collected by Keuleers, Stevens, Mandera, & Brysbaert (2015). To operationalize affect, we use concept polarity (the sensitivity of each concept to negative or positive affect) and the proportion of diminutive responses per concept.

The response variable takes two aspects of lexical geographical variation into account. On the one hand, we calculate the diversity of a concept as the number of word types that occurs in the dataset per concept. On the other hand, we also take into account geographical scatter. More specifically, we model the degree to which each concept and the variants per concept are scattered in a heterogeneous way across geographical space. We use (the logarithm of) the product of diversity and scatter as the response variable of our linear regression analysis.

IV. RESULTS

The linear regression model reveals that differences between semantic fields occur. More specifically, semantic fields that are more abstract (for instance the semantic field containing concepts that relate to a person’s feelings and personality) show more lexical variation than more concrete semantic fields (such as the semantic field of the human body). We are also able to compare the operationalizations of our predictor variables. In sum, our analysis provides further evidence for
the influence of semantic concept features on lexical geographical variation.

REFERENCES


