SCIENTIFIC REPORT

Research Networking Programmes (RNPs)
ESF Short Visit Grants

'European Network on Word Structure: Cross-disciplinary Approaches to Understanding Word Structure in the Languages of Europe' (NetWordS)

Project: "Meaning structure in emotion words"
Reference number: 4778

Grantee (visiting scholar): Dr. Cristina Soriano Salinas (University of Geneva, Switzerland)
Host: Dr. Dylan Glynn (University of Lund, Sweden)

Purpose of the visit

The goal of the short stay was to bring together two teams of researchers from Switzerland and Sweden to learn about each other's approach to lexical semantics and design a collaborative research line on the emotion lexicon. The disciplines involved are psychology, cognitive linguistics and corpus lexicology. We aimed to compare the insight provided by direct elicitation (self-report cued by questionnaires) and language observation (corpus-based analysis of lexeme use) in the study of the meaning of emotion words.

Description of the work carried out during the visit

The short grant allowed two teams to meet for 10 days and introduce each other to their respective theoretical frameworks, methodological techniques and findings on emotion word semantics.

The group at the Swiss Center for Affective Sciences (University of Geneva), represented by the applicant, brings together psychology and cognitive linguistics. In our past work, we have identified the semantic profiles of key emotion words across several languages and cultures on the grounds of web-questionnaires based on psychological theories of emotion, which capture native speakers' intuitions on the meaning of words (GRID and ELIN projects) (Fontaine, Scherer & Soriano, in press). This allows us to investigate universality/specificity, and the relative impact of language vs. culture on emotion word meaning by comparing languages spoken in the same country and countries that speak the same language. It also allows us to identify what aspects may be more relevant in the definition of specific lexemes in a language or cross-lingually.

During the short visit, the applicant presented this research line to the host team in a 2.5 h long public seminar. The discussion following the seminar proved very useful to identify areas of convergence between the approaches endorsed by each team, as well as aspects of the research where the Lund corpus-based methodologies could be of particular value.

The team led by the host researcher specializes in corpus lexicology. Their goal is the corpus-based investigation of key cultural concepts across and within languages on the grounds of lexical usage patterns (Feature Configuration Analysis, Geeraerts *et al* 1994; Profile-Based Analysis, Gries 2003). This involves the manual analysis of a wide range of linguistic and socio-cultural features for a large number of examples of natural uses of key

lexemes associated to the concept under investigation. The outcome is a large database of usage patterns that can be submitted to advanced multivariate statistics to obtain a multidimensional picture of how those lexemes are used.

During the short visit, the host scholar introduced the visitor to the basic methodological principles of the approach, illustrating them with case studies on emotion lexemes carried out in the framework of his teaching at the University of Lund. The instruction comprised several aspects: introduction to the corpora used by the host, learning how to use the corpus-search software, learning how to compose an annotated usage database, learning statistical methods of analysis for the usage database, learning how to carry them out in the platform R, and practicing the annotation and analysis technique with a test sample.

Once both teams were well acquainted with the methodologies employed by the other group, we set out to pool together our respective knowledge in order to design a semantic coding schema to investigate the lexical representations of the emotion "anger" in English. First, we selected 2 lexemes (*anger* and *irritation*) for which the visiting team had already acquired a semantic profile based on elicitation techniques (notice that one of the ultimate goals of the study was to compare the results of both approaches). Second, we decided to extend the target words to include the adjectival forms of those lemmas, since the meaning of nominal and adjectival forms of the same root have been shown to differ (Glynn 2007, 2009, 2010a). Third, the insight provided by the GRID study (and the various psychology theories represented in it) was used to select features relevant for the characterization of the lexemes in English, both in the US and the UK. These features were chosen for their saliency (for both language variants) in the 4 underlying semantic dimensions identified by GRID for the emotion lexicon: axiology/valence, power/potency, arousal/activation and unpredictability/novelty (Fontaine et al, 2007). Fourth, these features were complemented with grammatical and semantic ones typically used in cognitive linguistics (including, for example, whether the lexeme was used in a figurative expression and, in that case, what the source domain of the metaphor was). Our pilot coding schema included over 30 categories reflecting the lexico-grammatical characteristics of the words in use, and information on the emotion-eliciting situation, the emotional state itself, and the involved participants (i.e. the arguments in the semantic structure of the examples). Our next step was the manual annotation, according to the coding schema, of a random sample of 73 occurrences of the words *anger*, *angry*, *irritated* and *irritation* in the corpus of teenager on-line personal diaries ('blog' postings) property of the host researcher (Speelman & Glynn 2005). The last step was to apply the learned techniques in a pilot exploration of the differences between the 4 lexemes by means of correspondence analysis, cluster analysis and logistic regression.

Description of the main results obtained

As specified above, the main goal of the visit was to become acquainted with each other's methodologies and theoretical backgrounds. In the spirit of "learning by doing" we decided to run a pilot study on the lexicalization of "anger" in English. The main practical goals for the study were to learn the research methodology and to enrich the coding technique with principles emerging from emotion psychology. These goals were fulfilled. In addition, the exploratory exercise allowed us to identify three main variables sufficient to tease apart the meaning of "anger" vs "irritation" and of their respective nominal and adjectival realizations.

The database was first subjected to cluster analysis, revealing that, as expected, lemma was a better predictor of semantic similarity than part of speech (e.g., *irritation* and *irritated* are more similar than *anger* and *irritation*).

A logistic regression allowed us to identify the major variables responsible for the meaning differentiation between the two emotions. The logistic regression on the lexical distinction ('irritation' vs. 'anger'), at the level of lemma, produced a relatively strong model. Model selection was based on predictor significance alone and not predictive strength. Input models were restricted to semantic features comparable to those in the GRID questionnaire (Fontaine, Scherer & Soriano, in press). Although a range of variables were found to be significant, the most parsimonious model included just two variables, Intensity and Initiative to act. "Intensity" refers to the strength of the feeling, as illustrated by the presence of intensifiers (*very*, *extremely*), expletives (*fucking*), capitalization (*SHUT UP*) or exclamation marks. We distinguished three levels: low intensity feeling (Intensity.Non), intense feeling (Intensity.Medium) and very intense feeling (Intensity.Strong). "Initiative to act" refers to whether the person feeling the emotion wanted to act on it, for example seeking retribution, punishment or correction. We distinguished between cases where this desire to act was explicitly stated in the sentence (Emoter_Initiative. Explicit), implicit in the situation (Emoter_Initiative.Implicit) or cases where it was explicitly stated that there was *no* desire to do anything, but simply complain and resign (Emoter_Initiative.Non). Specifically, two features, Intensity.Non and Emoter_Initiative.Non were found to predict 'irritation'.

```
Logistic Regression Model
glm(formula = Lemma ~ Intensity + Initiative_to_act, family = "binomial", data = dataframe)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9233  -0.7289  -0.5271   0.9626   1.8934

Coefficients:
                                 Estimate    Std. Error z value Pr(>|z|)
(Intercept)                      -1.6103     0.8462    -1.903   0.0570 .
Intensity.Strong                 -0.2933     0.6098    -0.481   0.6305
Intensity.Non                    1.7187      0.8039    2.138    0.0325 *
Emoter_Initiative.Implicit       0.4204      0.9658    0.435    0.6634
Emoter_Initiative.Non            1.5700      0.8642    1.817    0.0692 .
EmInitiativeUnkn                 0.3352      1.3637    0.246    0.8059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

In order to determine the predictive strength of the model, a R2 (Nagelkerke pseudo R2 statistic) and a ROC were calculated. The pseudo R2 is a stable 0.311 (0.3 being considered the minimal score required for a stable model, Speelman 2012). The ROC (C) statistic is a logit based score that calculates the predictive accuracy of the model. Although not a probability score, it can be interpreted as an approximation of the percentage of correctly determined occurrences (Glynn 2010b). A score of 0.773 is not as high as one would expect with a R2 of 0.311. Although not an unpredictive model, the rule of thumb for a strongly explantory model is taken to be 0.8 (.5 is pure chance and .8 is 30% better than chance).

```
Logistic Regression Statistics

lrm(formula = Lemma ~ Intensity + Initiative_to_act, data = dataframe, x = T, y = T)

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes          Indexes
Obs             73    LR chi2     19.06    R2        0.311    C        0.773
 ANGER          44    d.f.            5    g         1.320    Dxy      0.545
 IRRITATION     29    Pr(> chi2) 0.0019    gr        3.742    gamma    0.599
max |deriv| 7e-08                          gp        0.269    tau-a    0.265
                                           Brier     0.182
```

Two of the possible explanations for the lack of predictive strength in the model include:

there is not a great difference in use between the two lemmata or a hidden variable is making the distinction less clear-cut than the model assumes. Since there is no reason to assume that the different grammatical profilings of the each lemma behave in the same way, it is very likely that at least one of the grammatical profilings lies between the two lemmata in usage. Due to the data sparseness inherent to the pilot study, it is not possible to run a model of the different lexemes. Instead, we can turn to an exploratory technique in order to see if there are reasons to believe that grammatical profilings represent a hidden variable the absence of which would affect the explanatory power of the model.

A correspondence analysis was therefore carried out with the identified variables to explore the relative position of the lexemes with respect to one another in this multidimensional space. In order to ascertain which factors could be interacting with the lexical behavior, a range of features found to be significant in the model selection process were included. The most intuitively coherent results (correspondence analysis is an exploratory technique designed to aid in the identification of interactions in categorical data, see Glynn 2012) were obtained by dividing the lemma into its different parts of speech and by adding the factors of Intensity, Initiative to act and Intentionality. "Intentionality" refers to whether the action motivating the emergence of anger (i.e. the "cause"of anger) was intentional. We distinguished between cases where there was explicit evidence in the sentence that the anger-eliciting action had been intentional (Cause_Intentionality.Present), cases where it was accidental (Cause_Intentionality.Non), and cases where there was no information in this respect (Cause_Intentionality.N.A).

As Fig. 1 illustrates, the 'irritation' lexemes differ from the 'anger' ones in the lower intensity of the experiences they designate (non intense or medium), and because the desire of the emoter to act on account of his/her anger is either inexistent or only implicit in the sentences. On the contrary, in the sentences with 'anger' lexemes the desire to act is more likely to be explicitly stated and the intensity of the experience is either medium or strong.

Differences can be found inside each emotional category as well. *Irritation* and *irritated* are different in that *irritated* is more linked to cases where the intentionality of the cause is explicit, whereas in *irritation* the act that elicits the emotion is more likely not be intentional. For the 'anger' terms the main difference lies in the intensity of the emotion: medium for *angry* and strong for *anger*.

Overall, *anger* and *angry* seem to be closer in meaning to one another than *irritation* and *irritated*. The closest resemblance between emotion categories is found between the noun *irritation* and the adjective *angry.* In other words, the closest term to *irritation* is not *anger*, but *angry*. Neither *anger* nor *angry* are close in meaning to *irritated.* This would explain the relative predictive weakness of the logistic regression model. More data will permit us to run the model on individual lexemes, but also to run a log-linear analysis (Glynn & Krawczak, submitted) which will enable confirmatory analysis of multinomial predictor variables such as a part-of-speech sensitive lemma (*anger*, *angry*, *irritated* and *irritation*).

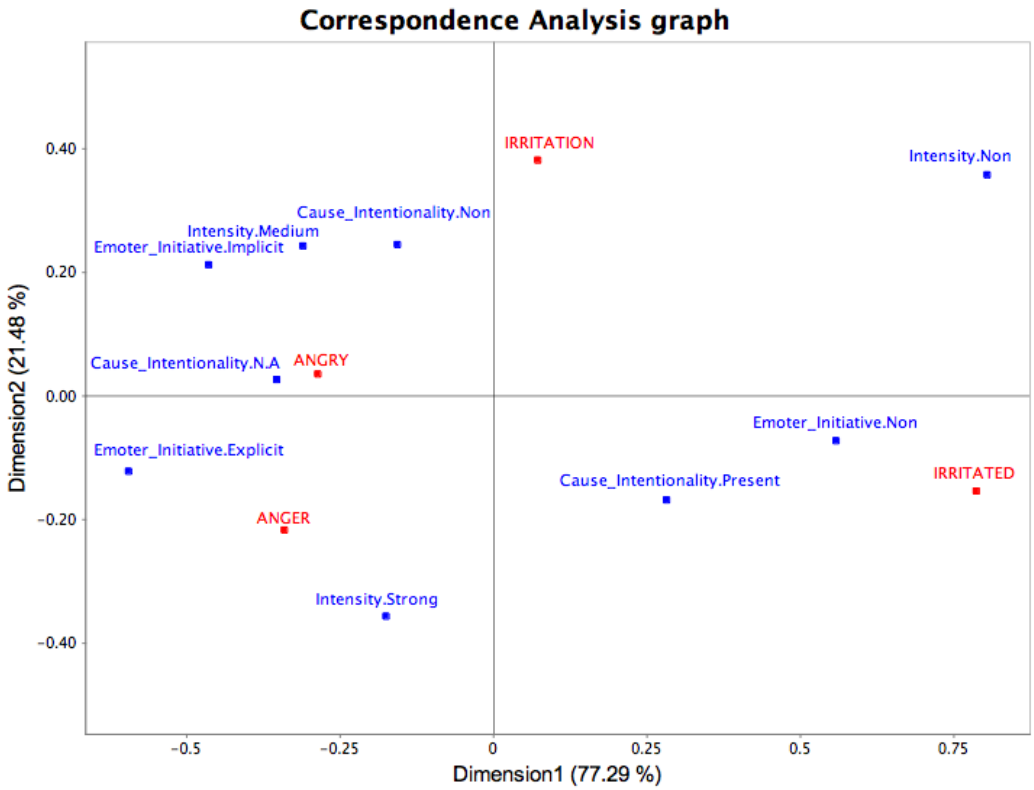

Fig. 1. Binary correspondence analysis, Lexeme, Intensity, Initiative and Intentionality.

Future collaboration with host institution

The two teams agreed to collaborate in the framework of a larger research project on social emotions to be submitted by the visiting researcher to the Swiss National Foundation next year. The goal is to investigate the same emotion concepts using elicited data, lexico-grammatical observation and metaphor analysis, as a way to triangulate findings and provide complementary insight where any one of the methods falls short.

Projected publications / articles to result from the grant

The two teams will work collaboratively on two papers on the emotion “anger” in English. One publication will target a psychology audience and the other a linguistic audience. In these papers the goal will be to compare the insight provided by both methods and raise awareness on the aspects that can be better tapped on by one or the other approach, as well as those semantic aspects that are typically overseen by one of them.

Other comments

Work on these papers will start in the fall semester 2012.

References

Fontaine, Johnny, Scherer, Klaus, Roesch, Etienne & Ellsworth, Phoebe (2007). The world of emotions is not two-dimensional. *Psychological Science, 18* (12): 1050-1057.

Fontaine, Johnny, Scherer, Klaus & Soriano, Cristina (Eds.) (in press). *Components of Emotional Meaning: A Sourcebook*. Oxford: OUP

Geeraerts, Dirk, Grondelaers, Stefan & Bakema, Peter. 1994. *The Structure of Lexical Variation. Meaning, naming, and context*. Berlin: Mouton de Gruyter.

Glynn, Dylan. 2007. "Mapping Meaning. Towards a usage-based methodology for Cognitive Semantics". PhD thesis, University of Leuven.

Glynn, Dylan. 2009. Polysemy, Syntax, and Variation. A usage-based method for Cognitive Semantics. V. Evans & S. Pourcel (eds.), *New Directions in Cognitive Linguistics*, 77-106. Amsterdam: John Benjamins.

Glynn, Dylan. 2010a. Synonymy, Lexical Fields, and Grammatical Constructions. A study in usage-based Cognitive Semantics. H.-J. Schmid & S. Handl (eds.), *Cognitive Foundations of Linguistic Usage-Patterns*, 89-118. Berlin: Mouton.

Glynn, Dylan. 2010b. Testing the Hypothesis. Objectivity and Verification in Usage-Based Cognitive Semantics. D. Glynn & K. Fischer (eds.), *Quantitative Cognitive Semantics. Corpus-driven approaches*, 239-270. Berlin: Mouton de Gruyter.

Glynn, Dylan. 2012. Correspondence Analysis. An exploratory technique for identifying usage patterns. D. Glynn & J. Robinson (eds.), *Polysemy and Synonymy. Corpus methods and applications in Cognitive Linguistics*, 133-180. Amsterdam: John Benjamins.

Glynn, Dylan. & Krawczak, Karolina. Social cognition, Cognitive Grammar and corpora. A multifactorial approach to subjectification. *Cognitive Linguistics*.

Gries, Stefan Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A study of particle placement*. London: Continuum Press.

Speelman, Dirk & Glynn, Dylan. 2005. LiveJournal Diary Corpus of British and American English. QLVL, Univeristy of Leuven.

Speelman, Dirk. 2012. Logistic regression. A confirmatory technique for variant comparison in corpus linguistics. D. Glynn & J. Robinson (eds.), Polysemy and Synonymy. *Corpus methods and applications in Cognitive Linguistics*, 181-230. Amsterdam: John Benjamins.