Proceedings of the LSA

Modeling the semantics of emotion: A culturally grounded approach to Flemish narratives

Ratna Kandala, Katie Hoemann *

Abstract. Expression of emotions through natural language is deeply embedded in culturally specific contexts and extends beyond simple lexical labels. A central difficulty lies in extracting structured semantic knowledge from unstructured daily narratives, a task particularly challenging for under-resourced language varieties such as Flemish (Belgian Dutch), which have historically received minimal computational attention. This study evaluates transformer-based models, specifically BERTopic, against traditional co-occurrence-based models (LDA) and clustering baselines (KMeans) on a uniquely large corpus of 24,854 daily narratives collected from 102 Dutch speakers over 70 days in Belgium, using both automated coherence metrics and a human evaluation. Our findings demonstrate that moving beyond frequency-based models is essential for semantically and culturally accurate analysis of naturalistic emotional language in under-resourced varieties.

Keywords. Flemish, daily narratives, emotion words, topic modeling, large language models, embeddings, under-resourced languages

1. INTRODUCTION. The automated extraction of structured meaning from natural language remains a central objective in computational linguistics, yet it faces a recurring obstacle: the inherent fluidity of human expression. Modeling the semantics of emotion presents a particular challenge in this regard, as emotional expression is deeply embedded in culturally specific contexts and extends far beyond simple lexical labels such as *happy* or *sad* (Wierzbicka 1999; Barrett 2017; Fontaine et al. 2013). The relationship between a word and its meaning(s) are not fixed but context-dependent and community-specific, a fact long recognized in linguistics (Firth 1957; Lyons 1977; Geeraerts 2010) but only partially addressed by computational methods. A central difficulty arises from extracting structured semantic knowledge from nuanced, unstructured daily narratives: the very genre in which emotional life is most richly and candidly expressed. This task is compounded in under-resourced languages and regional varieties such as Belgian Dutch (Flemish), which have historically received minimal computational attention compared to languages such as English or Spanish ((Jørgensen et al. 2015; Rehm & Uszkoreit 2012)), despite exhibiting rich sociolinguistic variation that standard Natural Language Processing (NLP) pipelines are ill-equipped to handle ((Grondelaers et al. 2011; Jaspers & Van Hoof 2015)).

This work investigates how transformer-based methods can capture the semantic specificity inherent in individual linguistic communities, with a focus on Flemish, a regional variety of Dutch. We evaluate the effectiveness of contextual embeddings derived from large language models, specifically BERTopic (Grootendorst 2022), against traditional co-occurrence-based models such as Latent Dirichlet Allocation Blei et al. (2003); Shin et al. (2025) and geometric clustering techniques such as KMeans (Berkhin 2006; Sinaga & Yang 2020). The evaluation is conducted on a uniquely large corpus of 24,854 daily narratives collected from 102 native Dutch speakers over 70 days in Belgium (Mestdagh et al. 2023).

* Authors: Ratna Kandala, University of Kansas (n038k926@ku.edu) & Katie Hoemann, University of Kansas (hoemann@ku.edu)

1.1. SOCIOLINGUISTIC CONTEXT: FLEMISH AS A CASE. Flemish is spoken in the northern region of Belgium (Flanders) and differs significantly from Netherlandic Dutch by displaying substantially more dialectal features, a richer inventory of Belgium-specific lexical items, and distinct pragmatic norms (Kestemont et al. 2012; Vandekerckhove & Nobels 2010; Grondelaers et al. 2011; Geeraerts et al. 1999). The sociolinguistic situation of Flemish is characterized by a register continuum ranging from regional dialects through *tussentaal* ('in-between language,' an informal colloquial variety not reducible to either dialect or standard Dutch) to standard Belgian Dutch (Jaspers & Van Hoof 2015; Ghyselen & Van Keymeulen 2016; Lybaert 2017). This continuum means that the texts of Flemish speakers produce in naturalistic settings are not well represented in standard Dutch corpora, and NLP tools calibrated to formal or Netherlandic text may systematically mishandle the lexical and morphological properties of everyday Flemish writing (Plank 2016). The corpus analyzed in this study differs importantly from the text types typically used in computational linguistics. Unlike news corpora, tweets, or product reviews, experience-sampling narratives are produced under mild ecological pressure, in informal registers, and with explicit affective framing (Csikszentmihalyi & Larson 1987; Hoemann et al. 2020).

2. PARTICIPANTS. 102 Dutch-speaking adults (N = 102, age M = 26.47, SD = 8.87, 52 women, 49 men, 1 other) residing in Belgium with a smartphone participated in this 70-day study. They were prompted to answer the following question four times a day: "What is happening right now and how do you feel about it?" They received the prompts via the m-Path Mestdagh et al. (2023) app on their smartphones. They responded either as typed text or recorded a voice message (which was transcribed later). The study was approved by KU Leuven Social and Societal Ethics Committee (protocol G-2023-6379-R3). This resulted in a corpus of around 24,752 texts in colloquial Flemish after filtering out texts that have fewer than 15 words. A corpus-specific stop word list was prepared and applied to reduce high-frequency, uninformative tokens (see Kandala et al. (2025a,b) for more details about data collection). To address the high lexical variability inherent in the Flemish narratives, we applied lemmatization using the Stanza pipeline (Qi et al. (2020)). This process collapses inflectional variants into a single base form, effectively reducing the sparse data problem while preserving the underlying semantic intent of the narratives. By grouping conjugated forms, we ensured that the models captured stable experiential contexts rather than grammatical noise.

3. MODELING. Before modeling, a corpus-specific list of emotion words was prepared by the authors. We subsequently masked these emotion words within the narratives before processing. This was done to ensure the models clustered the texts based on the experiential context surrounding the emotion rather than the lexical label of the emotion itself.
We compare three distinct approaches to topic modeling: First, Latent Dirichlet Allocation (LDA): A generative statistical model that assumes documents are a mixture of topics and topics are a mixture of words (Blei et al. 2003; Shin et al. 2025). While competitive in automated coherence, LDA treats language as a "bag of words," ignoring syntax and word order. Second, K-Means Clustering: A non-probabilistic method that partitions observations into k clusters based on vector distance. Third, Transformer-based approach BERTopic (Grootendorst (2022)) utilizes Self-Attention Mechanism to weigh the importance of all words in a sentence relative to one another (Vaswani et al. 2017). This allows us to move beyond simple one-to-one word mappings and uncover sophisticated, one-to-many mappings between emotion words and their varied contexts.

3.1. EVALUATION CRITERIA:. We assess model performance through two lenses: Quantitative: Using the C_v coherence metric, and Qualitative: Human evaluation of topic interpretability, focusing on whether the clusters yielded "culturally meaningful" insights into the Flemish experience.

4. RESULTS.

4.1. MODEL PERFORMANCE AND TOPIC COHERENCE. We first examine the quantitative performance of the three models. LDA achieved the highest coherence score of $C_v = 0.54$ compared to BERTopic ($C_v = 0.34$). However, a closer look at the outputs suggests a mismatch between automated metrics and semantic interpretability. BERTopic consistently produced coherent, culturally meaningful topics. For instance, BERTopic identified nuanced semantic clusters related to "studying," whereas LDA and KMeans produced diffuse, high-frequency terms lacking clear semantic cohesion. This advantage extended across various domains, including everyday routines (commuting by train) and leisure activities (watching television). In another case, when examining the topic of "fitness" (see Figure 1), we found that BERTopic (n=10 terms) produced a highly cohesive set including *fitnessen* ('to work out') and *kracht_training* ('strength training'). In contrast, LDA's cluster for the same domain was conflated with unrelated high-frequency terms such as *broer* ('brother') and *baby*. This indicates that for informal Flemish narratives, co-occurrence alone is an insufficient proxy for semantic relatedness.

To assess topic coherence as perceived by a language-competent human judges, we conducted a word intrusion task (Chang et al. 2009) in which two native Flemish-speaking annotators identified a semantically anomalous intruder word from sets of six topic-representative terms; BERTopic achieved markedly higher intruder detection accuracy (95.0%, 87.5%) than LDA (20.0%, 25.0%) or KMeans (60.0%, 52.5%), with substantially stronger inter-annotator agreement (Krippendorff's $\alpha = 0.874$) than either baseline ($\alpha \approx 0.547$) Krippendorff (2004), confirming that automated coherence scores systematically underestimate the semantic quality of embedding-based models (Lau et al. (2014)).

4.2. MAPPING THE FLEMISH EXPERIENTIAL LANDSCAPE. Next, we zoom in on the culturally specific clusters identified by the Transformer-based model, BERTopic. It successfully isolated topics that are semantically load-bearing in the Belgian context, such as:

- *Chiro*: Participation in Flanders largest youth organization.
- The "Rain" Cluster: A distinct semantic space for dissatisfaction related to the Belgian climate.
- Commuting: Detailed clusters related to *pendelen* 'commuting by train'.

Crucially, the model revealed *one-to-many mappings* for emotion words. The word *trots* 'proud' was not a monolithic label; rather, it was semantically anchored in three distinct contexts: (1) personal accomplishments, (2) work meetings, and (3) mundane routines like "walking the dog." This suggests that in Flemish daily life, "pride" is as much about the domestic and the routine as it is about professional achievement.

5. Discussion. The results of this study open onto three questions of broader significance for linguistics and the computational study of meaning.

BERTopic	LDA	KMeans
fitnessen (to work out), *fitness*, *workout*, *gym*, *oefening* (exercise), *trainen* (to train), *joggen* (to jog), *kracht_training* (strength training), *sport_les* (sports class), *sport_school* (sports school)	*fitnessen* (to work out), *broer* (brother), *baby*, *time*, *start*, *hasten* (to hurry), *toe_voegen* (to add), *quality*, *model*, *aanpassing* (adjustment)	*fitness*, *goed* (good), *eten* (to eat), *dag* (day), *vandaag* (today), *weten* (to know), *leuk* (nice/fun), *zin* (mood), *vanavond* (tonight), *moe* (tired), *beginnen* (to begin), *studeren* (to study), *rest*, *werken* (to work), *proberen* (to try)

Figure 1. Topics identified by each model related to *fitness*. Flemish words are italicised; English glosses appear in parentheses.

Paradigmatic coherence as the criterion for semantic quality. The dissociation between automated coherence scores and the human intruder task accuracy is not merely a methodological inconvenience: it exposes a theoretical mismatch at the heart of corpus-semantic evaluation. Automated metrics such as C_v operationalize coherence as lexical co-occurrence within document windows, a syntagmatic criterion (de Saussure 1916; Lyons 1977). Human judges, by contrast, evaluate whether a set of words constitutes a recognizable semantic field, a judgment grounded in paradigmatic knowledge: the network of substitutability relations, collocational expectations, and encyclopedic associations that a speaker acquires as a member of a linguistic community (Murphy 2003; Geeraerts 2010). This finding resonates with psycholinguistic evidence that human lexical access is organized paradigmatically (Miller & Charles 1991; Budanitsky & Hirst 2006) and with usage-based accounts that treat semantic knowledge as accumulated over contextually diverse instances rather than stable co-occurrence frequencies (Bybee 2010; Tomasello 2003).

What counts as noise is variety-relative. The underperformance of K-Means, despite its documented efficacy on Standard Dutch corpora (Kamiloğlu et al. 2025), illustrate a principle with implications well beyond this study: the notion of *noise* in text preprocessing is not a neutral technical category but a variety-relative one. Standard TF-IDF vocabulary pruning eliminates low-frequency tokens on the assumption that they are uninformative (Biber et al. 1998). In a Flemish experience-sampling corpus, however, low-frequency items include dialectal diminutives, compound lemmas such as *aula_vriend* ('lecture-hall friend'), and culture-specific proper nouns such as *Chiro*, precisely the vocabulary through which speakers name the social situations of their emotional lives. This is an instance of the broader problem that Eisenstein (2013) and Plank (2016) identify as non-canonical language: preprocessing pipelines designed for standard, formal varieties systematically erase the features that distinguish regional and informal speech communities. The implication for corpus linguistics and NLP is that stopword lists, frequency thresholds, and lemmatization strategies must be variety-calibrated, not imported wholesale from resources built for prestige varieties (Jørgensen et al. 2015).

Culturally situated meaning and the limits of language-general models. The emergence of *Chiro* and weather-related affect as stable BERTopic topics reflect a property of lexical mean-

ing that has been theorized in frame semantics (Fillmore 1985), cultural linguistics (Wierzbicka 1999; Goddard & Wierzbicka 2014), and the cross-linguistic study of emotion (Fontaine et al. 2013; Jackson et al. 2019): the semantic content of words is partly constituted by the culturally specific situations they are in conventionally used to describe. Language-general models, trained on multilingual corpora without community-specific calibration cannot recover this situated meaning because they average over the very variation that carries it. The present results add corpus-driven, quantitative evidence to what has previously been argued on theoretical and typological grounds: that lexical semantic descriptions must be anchored to specific speech communities if they are to be psychologically and culturally accurate (Barrett 2017; Mesquita & Frijda 1992).

6. Conclusion. This study set out to ask whether contextual embedding models can recover the culturally specific, paradigmatically organized semantic structure of emotion vocabulary from naturalistic daily narratives in an under-resourced variety. The answer is affirmative, but the route to that answer is as important as the conclusion itself. The decisive evidence came not from automated coherence metrics - which favored the traditional LDA baseline - but from native-speaker human evaluation, which revealed a near-perfect alignment between BERTopic's topics and the semantic fields recognized by Flemish annotators. This dissociation is itself a finding: it demonstrates that the evaluation framework matters as much as the model, and that automated metrics calibrated to co-occurrence statistics are inadequate judges of paradigmatic semantic quality.

Three empirical claims are established. Emotion words in naturalistic Flemish discourse participate in one-to-many mappings with experiential contexts: *trots* ('proud') is distributed across achievement, professional, and domestic frames; *verdrietig* ('sad') clusters more narrowly around loss and illness. These mappings are recoverable from corpus data through contextual embedding models without requiring language-specific pretraining, provided that embedding selection is validated qualitatively. And the notion of preprocessing noise is variety-relative: vocabulary items that frequency-based pipelines treat as uninformative - dialectal compounds, culture-specific proper nouns, - are precisely the semantic anchors of community-specific emotional meaning.

For the linguistics community, the broader contribution is methodological: a procedure for *experiential lexicography* (Hanks 2013), the empirical description of emotion words through the structured inventory of situations they characteristically name. This complements existing resources (dimensional affect ratings (Warriner et al. 2013), LIWC categories (Pennebaker et al. 2001), sentiment lexicons (Mohammad & Turney 2013)) by capturing situational specificity and cultural embeddedness rather than valence or arousal alone. Applied across languages and varieties, it offers a bottom-up, corpus-driven method for investigating the cross-linguistic semantic structure of emotion that neither elicited ratings nor lexical typology alone can provide (Fontaine et al. 2013; Jackson et al. 2019).

Future work will extend the emotion-word neighborhood analysis to the full affective vocabulary of the corpus, examine how context-mappings vary across speakers by age, gender, and social network position, and test whether the Flemish-specific semantic structures identified here diverge systematically from Netherlandic Dutch equivalents, a question with direct implications for theories of meaning variation across varieties (Geeraerts et al. 1999; Grondelaers et al. 2011) and for the construction of cross-culturally valid affective computing resources (Barrett 2017; Fontaine et al. 2013).

References

Barrett, Lisa Feldman. 2017. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Berkhin, Pavel. 2006. A survey of clustering data mining techniques. In Jacob Kogan, Charles Nicholas & Marc Teboulle (eds.), *Grouping multidimensional data*, 25–71. Springer.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use* Cambridge Approaches to Linguistics. Cambridge: Cambridge University Press.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3(Jan). 993–1022.

Budanitsky, Alexander & Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* 32(1). 13–47.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge University Press.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang & David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems 22 (neurips)*, .

Csikszentmihalyi, Mihaly & Reed Larson. 1987. Validity and reliability of the Experience-Sampling method. *Journal of Nervous and Mental Disease* 175(9). 526–536.

Eisenstein, Jacob. 2013. What to do about bad language on the internet. In *Proceedings of naacl-hlt*, 359–369.

Fillmore, Charles J. 1985. Frames and the semantics of understanding, vol. 6 2, 222–254.

Firth, John R. 1957. A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis* 1–32. Special volume of the Philological Society. Blackwell.

Fontaine, Johnny R. J., Klaus R. Scherer & Cristina Soriano (eds.). 2013. *Components of emotional meaning: A sourcebook*. Oxford University Press.

Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford University Press.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat*. Meertens Instituut.

Ghyselen, Anne-Sophie & Jacques Van Keymeulen. 2016. Tussentaal in Vlaanderen: een empirisch onderzoek naar definitie en frequentie. *Taal en Tongval* 68(2). 93–128.

Goddard, Cliff & Anna Wierzbicka. 2014. *Words and meanings: Lexical semantics across domains, languages, and cultures*. Oxford University Press.

Grondelaers, Stefan, Roeland van Hout & Pieter Muysken. 2011. How 'standard' is Standard Belgian Dutch? *Journal of Germanic Linguistics* 23(3). 243–268.

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* .

Hanks, Patrick. 2013. *Lexical analysis: Norms and exploitations*. MIT Press.

Hoemann, Katie, Fei Xu & Lisa Feldman Barrett. 2020. Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology* 56(3). 511–524.

Jackson, Joshua Conrad, Joseph Watts, Teague R. Henry et al. 2019. Emotion semantics show both cultural variation and universal structure. *Science* 366(6472). 1517–1522.

Jaspers, Jürgen & Sarah Van Hoof. 2015. Ceci n'est pas une tussentaal: Evoking standard and vernacular language through mixed Dutch in Flemish telecinematic discourse. *Journal of Germanic Linguistics* 27(1). 1–44.

Jørgensen, Anna, Dirk Hovy & Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text, acl*, 9–18.

Kamiloğlu, Roza G. et al. 2025. What makes us feel good? A data-driven investigation of positive emotion experience. *Emotion* 25(1). 271–276.

Kandala, Ratna, Niels Vanhasbroeck & Katie Hoemann. 2025a. Evaluating BERTopic on open-ended data: A case study with belgian dutch daily narratives. https://doi.org/10.48550/arXiv.2504.14707.

Kandala, Ratna, Niels Vanhasbroeck, Bastiaan Tamm, Hugo Van hamme, Peter Kuppens, Batja Gomes Gomes de Mesquita & Katie Hoemann. 2025b. Toward an ecology of emotion in everyday life. *PsyArXiv* https://doi.org/10.31234/osf.io/6f4q7_v1.

Kestemont, Mike, Walter Daelemans & Guy De Pauw. 2012. Robust rhymes? the stability of authorial style in Dutch literature. In *Proceedings of the eacl workshop on computational approaches to deception detection*, Cited for Flemish computational corpus work.

Krippendorff, Klaus. 2004. Content analysis: An introduction to its methodology, Sage.

Lau, Jey Han, David Newman & Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th conference of the european chapter of the association for computational linguistics (eacl)*, 530–539. Gothenburg, Sweden: Association for Computational Linguistics.

Lybaert, Chloé. 2017. Tussentaal in fictional dialogue: A corpus-based approach. *Dutch Journal of Applied Linguistics* 6(2). 216–237.

Lyons, John. 1977. *Semantics*, vol. 2. Cambridge University Press.

Mesquita, Batja & Nico H. Frijda. 1992. Cultural variations in emotions: A review. *Psychological Bulletin* 112(2). 179–204.

Mestdagh, Maarten, Stijn Verdonck, Maarten Piot, Kris Niemeijer, Ghina Kilani, Francis Tuerlinckx, Peter Kuppens & Eline Dejonckheere. 2023. m-path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. *Frontiers in Digital Health* 5.

Miller, George A. & Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1). 1–28.

Mohammad, Saif M. & Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29(3). 436–465.

Murphy, M. Lynne. 2003. *Semantic relations and the lexicon*. Cambridge University Press.

Pennebaker, James W., Martha E. Francis & Roger J. Booth. 2001. *Linguistic inquiry and word count: LIWC 2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

Plank, Barbara. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of konvens 2016*, .

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*, 101–108. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14.

Rehm, Georg & Hans Uszkoreit (eds.). 2012. *The language white paper series*. META-NET.
de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Payot. English trans. Wade Baskin, 1959.
Shin, Eunyong, Sun Yim & A Ra Koh. 2025. Comparison of consumer perceptions of sustainable and ethical fashions pre- and post-COVID-19 using LDA topic modeling. *Humanities and Social Sciences Communications* 12(1). 226.
Sinaga, Kusuma Prayoga & Ming-Shing Yang. 2020. Unsupervised k-means clustering algorithm. *IEEE Access* 8. 80716–80727.
Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
Vandekerckhove, Reinhild & Judith Nobels. 2010. Code eclecticism: Linguistic variation and code alternation in the chat language of Flemish teenagers. *Journal of Sociolinguistics* 14(5). 657–677.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30.
Warriner, Amy Beth, Victor Kuperman & Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4). 1191–1207.
Wierzbicka, Anna. 1999. *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press.