

The rating of texts for an International Database of Emotional Short Texts

First results

Egon Werlen & Per Bergamin

Institute for Research in Open-, Distance- and eLearning - IFeL

Swiss Distance University of Applied Sciences - FFHS

extended version



United Nations
Educational, Scientific and
Cultural Organization



UNESCO Chair on Personalized and
Adaptive Distance Education at the Swiss Distance
University of Applied Sciences, Brig, Switzerland



Abstract

The purpose of this exploratory study is to rate and categorize emotional characteristics of short texts (900-1100 characters) with the aim to create a balanced database for research.

Two different procedures were applied to evaluate valence and arousal of 24 German texts written by Swiss volunteers, 19 texts written by Catalan/Spanish and 35 texts written by French volunteers (with subsequent translation), and six further texts, drafted by Swiss researchers. Self-Assessment-Manikins (SAM) were used for an expert rating and the Berlin Affective Word List (BAWL) for an automated lexical rating. Texts from Switzerland, Spain and French show similar average values regarding valence and arousal. The correlation between SAM- and lexical-ratings for valence is $r=.61$, for arousal $r=.23$. The categorization of valence yields similar groups with SAM or lexical ratings. These collected text samples are not sufficient (e.g. overrepresentation of high arousal), therefore the additional drafting of stories with support from the lexical word ratings is recommendable.



Content

- Study
 - Goal of the study,
Research Questions
- Theory
- Methods
 - Text and Ratings
- Results
- Discussion
- Conclusion



Goal and Research Questions of the explorative Study



Goal of the study

The purpose of this exploratory study is to rate and categorize emotional characteristics of short texts (900-1100 characters) with the aim to create a balanced database for research.

Research Question

- How are the texts distributed regarding valence and arousal?
- Are there cultural differences in the rated texts?
 - Texts written by authors of different languages / cultures
- What's the relationship between automated lexical analysis and human ratings?



Theoretical background

■ Working model

- Components of User Experience Model (CUE model) by Mahlke and Thüring (2007)
- Model of working memory by Baddeley (Baddeley & Logie, 1999)
- Cognitive Theory of Multimedia Learning (CTML) by Meyer & Moreno (2003)
- Cognitive Affective Theory of Media Learning (CATLM) by Moreno (Moreno & Mayer, 2007; Moreno, 200)
- Cognitive Load Theory by Sweller (1999)
- Augmented Cognitive Load Theory (aCLT) by Huk & Ludwigs (2009)

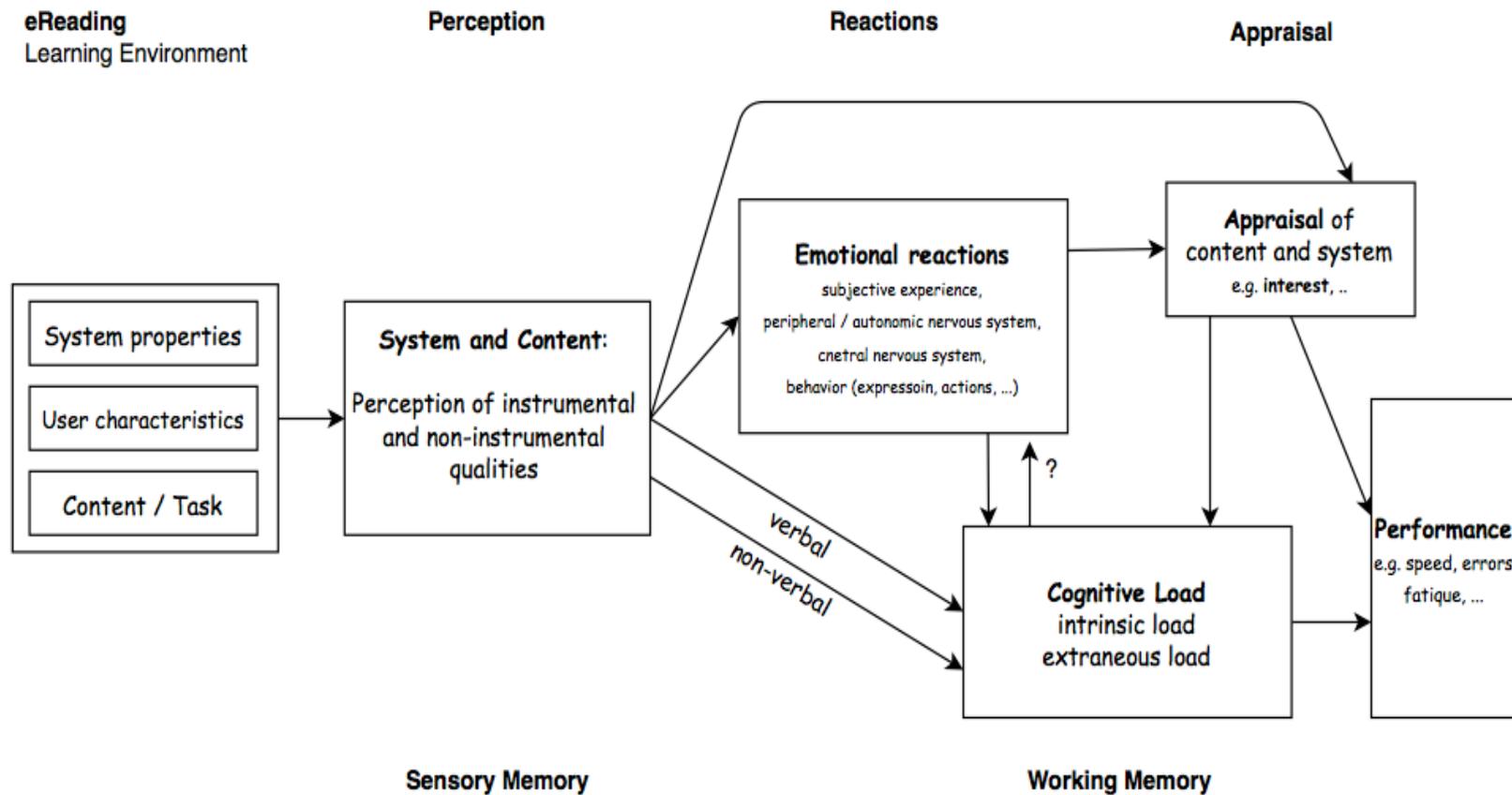
■ Alternative Model: **Neurocognitive Poetics Model**

(NCPM; Jacobs, 2015, Jacobs, Lüdke & Meyer-Sickendiek, 2013)

- Hsu et al. (2015): prediction of emotional potential by lexical and inter-lexical affective variables (using excerpts from Harry Potter novels)
- Ullrich et al. (2017): prediction of emotional potential by lexical, sublexical and inter-lexical affective variables (using 57 poems by H.M. Enzensberger)



Working model





Working Model

The model is based on the Components of User Experience Model (CUE model) by Mahlke and Thüring (2007), that was developed to illustrate the emotional experiences of learners with learning material. The CUE model consists of three components: perception of the instrumental qualities (closely related to the usability of a system, e.g. typeface, screen size, reading media), perception of the non-instrumental qualities (resulting from the incentive and attractiveness of the system, e.g. content, visual aesthetics, haptic quality) and the emotional reactions due to these perceptions (subjective feelings, motor expression, and physiological reactions). The reader's experiences that trigger the emotional reactions occur in interaction with the system, which always takes place in a certain context, over a certain time and within a certain text. Three components lead to the evaluation of the material. Depending on the evaluation of the reading/learning material and the overall situation, different emotions result that promote or inhibit the reading process. In addition to the components of the CUE model, the limited capacity of working memory (e.g. Baddeley & Logie, 1999) plays a role. The Cognitive Theory of Multimedia Learning (CTML) by Meyer & Moreno (2003) proposes two independent channels (verbal, visual) from perception, over working to the long-term memory. Moreno (Moreno & Mayer, 2007; Moreno, 2006) extended this model to the Cognitive Affective Theory of Media Learning (CATLM), in which the influence of emotions (and other affective variables) on the different channels of working memory (visual, non-visual) and thus to learning is involved. The understanding and recollection of a text largely correspond to a learning process. The Augmented Cognitive Load Theory (aCLT, Huk & Ludwigs, 2009, based on the Cognitive Load Theory by Sweller (1999), divides the cognitive load of the working memory in intrinsic load (caused by the nature of the content, e.g. difficulty of the task), extraneous load (representation of the text, i.e. unnecessary information for learning) and the learning load (germane load; necessary for schema mapping). Huk et al. (2009) expanded the theory by assuming that the learning-related load can be altered by cognitive interventions that help to construct mental models and affective interventions that support interest. The evaluation of a text is a process in which an event is judged in terms of valence, power, arousal, and novelty (see Ellsworth & Scherer, 2003). The primary appraisal is a rapid and automatic process. The secondary appraisal is a more slow and conscious process. The emotional appraisal is a continuous process, and the original appraisal can be modified by new experiences (reappraisal), while conscious cognitive processes also play a role.

Study – Theory – Methods – Results – Discussion – Conclusion



Neurocognitive Poetics Model

The Neurocognitive Poetics Model (NCPM; Jacobs, 2015, Jacobs, Lüdke & Meyer-Sickendiek, 2013) states that the three general factors reader, text, and context determine the mental responses and behavioural responses to literary texts. The central hypothesis with regard to the text states that background characteristics of the text (BG) and foreground properties of the text (FG) at least partially activate different neural networks and cognitive-affective processes. Background features are "elements of a text that create a feeling of familiarity in the reader" (p. 7) e.g. known words and propositions, which, analogous to gestalt psychology, form the background which the foreground properties stand out from. FG are the parts of a text which can be represented by parallelism or alienation (for example stylistic means, rhetorical figures, tropes, schemata). As a rule, the BG activate the fluid, mainly in the left hemisphere, of reading and immersion. This is where the fiction feelings hypothesis fits in, according to which "narratives with emotional contents invite readers more to be empathic with the protagonists and immersed in the text world [...] than do stories with neutral contents" (p. 3f). These processes are supported by affective systems as described by Panksepp (1998, 2012) in his theory of emotions. Panksepp (1998) assumes that emotional processes in reading are neurologically formed in the same systems in which all the other emotional processes run that we share with all mammals (Jacobson-Panksepp hypothesis). This means that nature did not have time to develop new brain structures, and the emotional processes take place in the 'old' brain structures. In the case of the BG, this affects, in particular, the systems of anxiety and anger, which are involved in feelings which arise during reading and which cause tension to build up. The FG properties generally result in a slower reading and facilitate aesthetic emotions (e.g., curiosity, pleasure, pleasure, sadness) that are increasingly processed in the right hemisphere. These processes are supported by the systems lust and play. The corresponding neurological processes, however, are not so clearly distinguishable and localizable, but overlap. Thus, the system can not only facilitate aesthetic emotions, but also participate in pleasant processes of immersion (Jacobs, 2015; Jacobs et al., 2013).



Texts and Ratings

■ Texts

- 24 texts by Swiss volunteers
- 19 texts by Spanish volunteers translated from Catalan/Spanish
- **new:** 35 texts by French volunteers translated from French (included)

■ Human raters

- 5-6 experts (collaborators at the IFeL)

■ Human ratings

- Self-Assessment-Manikins (SAM) for valence (-4 through 4) and arousal (1 through 9)
- Subjective rating of comprehensibility (1 through 9)

■ Automated lexical ratings

- Berlin Affective Word List (BAWL) for valence (-3 through 3) and arousal (1 through 5)
- Flesch-Index for readability (0 through 100)
- DeReWo (Deutscher Referenz Korpus) for text difficulty (word frequency; standardized 0-1):
<http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html>



Interrater reliability - ICC

	Valence	Arousal	Comprehensibility
■ ICC (5 raters) of Swiss, Catalan, French texts			
□ single measure:	.53	.13	.27
□ average measure:	.85 (α)	.43 (α)	.65 (α)

Single measures ICC: reliability if the amount of raters is reduced to one

Average measures ICC: agreement between the raters

Method: random, consistency (5 raters, 84 texts, incl. 6 constructed texts)

- The automated lexical ratings (BAWL-R, i.e. lists of rated words) have a theoretical reliability of 1.00



Results – Categories

SAM-ratings valence negative valence positive

arousal low	13	16.7%	20	25.6%
arousal high	31	39.7%	14	17.9%

split: centre of scale

BAWL-R valence negative valence positive

arousal low	11	14.1%	28	35.9%
arousal high	28	35.9%	11	14.1%

split: median

- The categorization with SAM- and lexical ratings yields similar results.
- The collected text samples overrepresent negative valence / high arousal and positive valence / low arousal



Results - Averages

Origin	SAM		BAWL-R (automated)					
	valence	arousal	valence	arousal	valence	arousal		
Switzerland	4.86	(1.35)	5.48	(0.93)	0.66	(0.28)	2.63	(0.12)
Catalan	4.94	(1.61)	4.80	(0.97)	0.67	(0.22)	2.65	(0.09)
France	4.47	(1.57)	5.08	(1.09)	0.63	(0.23)	2.63	(0.11)

	SAM		BAWL-R	
	valence	arousal	valence	arousal
F	0.78	2.45	0.17	0.36
p	.464	.094	.844	.701

Texts from Switzerland, Spain, and French show similar average values regarding valence and arousal both with SAM-ratings and lexical ratings (BAWL-R).



Results – Averages

Origin	rating		automated		N
	comprehensibility	Flesch	difficulty		
Switzerland	7.20 (0.92)	70.04 (10.52)	0.63 (0.18)	24	
Catalan	7.03 (0.60)	63.21 (7.50)	0.46 (0.14)	19	
France	6.76 (0.73)	63.69 (12.09)	0.50 (0.18)	35	

	rating	automated	
	compreh.	Flesch	difficulty
F	2.45	3.12	6.08
p	.093	.050	.004

There are differences between the texts concerning readability (Flesch Index) and text difficulty (DeReWo frequency), but not regarding person-rated comprehensibility.



Results - Correlations

- Swiss, Catalan, and French texts:
Correlations between SAM-ratings and automated
lexical-ratings:

valence **$r=.61$** ($p=.000$)

arousal **$r=.23$** ($p=.045$)



Discussion

- There is a growing amount of research in automated detection of emotions in texts. There are different methods used (ontology / keyword based, machine learning, and mixed approaches). Strapparava et al. (2010) analysed several methods. The highest correlation between automated detected emotions and human ratings for valence was **$r=.48$** (knowledge-based domain-unspecific non-supervised approach – CLaC)
- Authors from the research group of Arthur Jacobs (Berlin) used automated lexical ratings. Ullrich et al. (2017) had 252 students rate the German poems by H.M. Enzensberger's "Die Verteidigung der Wölfe" in terms of valence and arousal. A combination of lexical, interlexical, and sublexical characteristics of the poems predicted the rated valence with 59%, the arousal with 43%. The correlation of lexical (BAWL-R) with rated valence was **$r=.65$** , of lexical with rated arousal **$r=.54$** .
- Hsu et al (2015) analysed 120 short excerpts from the Harry Potter novels (in German). 24 students read the texts in a fMRI scanner and rated them afterwards subjectively on valence and arousal. The correlation of lexical (BAWL-R) with rated valence was **$r=.53$** , of lexical with rated arousal **$r=.59$** .



Discussion

The correlations of automated lexical and subjectively rated emotional valence are similar to the results found in comparable studies. The value for emotional arousal is much lower in our study. That may be due to the low interrater reliability.

Valence

- This study: **r=.61**
- Best value of study by Strapparava et al. (2010): **r=.48**
- Poems (by Enzensberger); Ullrich et al. (2017): **r=.65**
- Novel excerpts (Harry Potter); Hsu et al. (2015): **r=.53**

Arousal

- This study: **r=.23**
- Poems (by Enzensberger); Ullrich et al. (2017): **r=.54**
- Novel excerpts (Harry Potter); Hsu et al. (2015): **r=.59**



Conclusion

- For valence, automated lexical analyses correspond to a certain degree with person-made ratings
- Arousal seems to be a more complicated issue
 - We need further studies (other raters, other instruments: SAM vs. questions)
- Advantages of automated lexical analyses:
repeatable, reliable, independent of rater-sample, faster
- -> Automated lexical analyses of texts including learning material will be usable in the near future

- **Next steps:**
 - Rating of 60 texts (Switzerland*, Germany) and 3 neutral texts by at least 20 students
 - Translation of all (CHE, DEU, FRA, ESP, FIN, TUR, PRT, GRC) texts in English
 - Rating of English texts
 - Translation in languages from participating countries

* 3 texts from Catalan / Spain; 3 texts from France



Questions? / Comments!

I can see Sabine. She lies on her back. In a distance, she hears steps, behind the wooden fence something is rustling. A silent breeze touches her face. Her foot is in pain. Sabine can't see it under the grass. The shoe is still hanging on the fence. When I see her face, I have to hold myself. My head's dizzy. Sabine is pale; she's white in her face. She should be afraid but she looks quietly at me with a small smiling behind the aching. The steps are approaching. I don't know what to do. A little snake winds through the fence and disappears in the high grass. The steps are getting louder. It has to be behind the garage. The light above the forecourt flashes. It gets silent. The wind passes gently and silently over us. There is this scratching behind the garage. A piece of metal snaps aloud. Sabine begins to yell. Suddenly, the two men are standing there. The stretcher snaps with a click during positioning.

https://www.researchgate.net/profile/Egon_Werlen/contributions



References

- Baddeley, A.D. & Logie, R.H. (1999). Working memory. The Multiple-Component Model. In A. Miyake & P. Shah (Hrsg.). *Models of working memory: Mechanisms of active maintenance and executive control* (S. 28-61). Cambridge: University Press.
- Huk, T., & Ludwigs, S. (2009). Combining cognitive and affective support in order to promote learning. *Learning and Instruction*, 19(6), 495-505.
- Hsu, C. T., Jacobs, A. M., Citron, F. M. & Conrad, M. (2015). The emotion potential of words and passages in reading Harry Potter—An fMRI study. *Brain and language*, 142, 96-114.
- Jacobs, A.M. (2015). Neurocognitive poetics: methods and models for investigating the neuronal and cognitive-affective bases of literature reception. *Frontiers in Human Neuroscience*, 9:186. doi: 10.3389/fnhum.2015.00186
- Jacobs, A. M., Lüdtke, J., and Meyer-Sickendiek, B. (2013). Bausteine einer neurokognitiven poetik: foregrounding/backgrounding, lyrische stimmung und ästhetisches gefallen in B. Meyer-Sickendiek and F. Reents (Hrsg.). *Stimmung und Methode* (pp. 63–94). Tübingen: Mohr Siebeck.
- Mahlke, S. & Thüning, T. (2007). Studying Antecedents of Emotional Experiences in Interactive Contexts (p. 915-918). *CHI '07: Proceedings*. ACM Press.
- Meyer, R.E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43-52.
- Moreno, R. (2006). Does the modality principle hold for different media? A test of the method-affects-learning hypothesis. *Journal of Computer Assisted Learning*, 22(3), 149-158.
- Moreno, R., & Mayer, R. (2007). Interactive Multimodal Learning Environments Special Issue on Interactive Learning Environments: Contemporary Issues and Trends. *Educational Psychology Review*, 19(3), 309-326.
- Panksepp, J. (1998). *Affective Neuroscience*. New York: Oxford University Press.
- Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Camberwell: ACER Press.
- Ullrich S., Aryani A., Kraxenberger M., Jacobs, A.M. & Conrad, M.(2017) On the Relation between the General Affective Meaning and the Basic Sublexical, Lexical, and Inter-lexical Features of Poetic Texts—A Case Study Using 57 Poems of H. M. Enzensberger. *Frontiers in Psychology*, 7(2073). doi: 10.3389/fpsyg.2016.0207