Predicting Readers' Emotional States Induced by News Articles through Latent Semantic Analysis

Diana Lupan, Stefan Bobocescu-Kesikis, Mihai Dascalu, Stefan Trausan-Matu, Philippe Dessus

In B. Patrut (Ed.)(2013), 1st Int. Conf. Social Media in Academia: Research and Teaching (SMART 2013) (pp. 79–84). Bologna: Medimond-Monduzzi.

Predicting Readers' Emotional States Induced by News Articles through Latent Semantic Analysis

Diana Lupan
University Politehnica of Bucharest, Romania
diana.lupan@cti.pub.ro

Stefan Bobocescu-Kesikis
University Politehnica of Bucharest, Romania
stefan.bobocescu-kesikis@cti.pub.ro

Mihai Dascalu
University Politehnica of Bucharest, Romania
mihai.dascalu@cti.pub.ro

Stefan Trausan-Matu
University Politehnica of Bucharest, Romania
stefan.trausan@cs.pub.ro

Philippe Dessus
University Grenoble Alpes, LSE, France
philippe.dessus@upmf-grenoble.fr

Abstract:
With the increasing spread of the social web, identifying emotions in texts has proved to have various applications in fields like opinion mining or market analysis. Emotion recognition from written statements does not only reveal information about the person who wrote them, but can also be used in predicting how the emotional state of the readers can be affected. We propose a novel automatic method for analyzing texts that predicts how reading a news article can influence in turn the emotional state of the reader. This method integrates several word-count approaches and natural language processing techniques, such as Latent Semantic Analysis. Moreover, our implemented system contains a module designed to personalize the provided feedback according to the reader's current emotional state. A preliminary validation has been performed and results are promising.
Keywords: automatic evaluation of news articles, emotional state, Latent Semantic Analysis

1. Introduction

Emotions play an important role in our lives as they can be identified in various contexts (e.g., arguing with someone, reading an article, receiving or giving bad news). Usually, emotions can be easily identified when analyzing body language or voice features (tone, rhythm, frequency). On the contrary, a particular case is when none of the previous cues are valid, as the persons involved are not present face-to-face or not even able to communicate verbally, but only through written communication. We will focus on this situation and analyze how a news article can influence the emotional state of the person reading it.

The system we built, Emo2 (Emotions Monitor) [9], was implemented in two phases. Firstly, we enforced a context independent approach that predicted the emotional state by evaluating the content and the title of a news article using different natural language processing techniques. Afterwards, we added a second stage of analysis (results personalization), based on the current emotional state of the user who is reading the news articles. In comparison to the first release of the system [9], the number of considered articles increased dramatically and a multitude of fine-tunes were performed in order to increase the system's precision. Also, another validation was performed in order to verify the validity of the results.

The remainder of the paper presents similar work, the system's architecture and details about the actual evaluation steps. The last two sections are focused on the presentation of results and conclusions.

2. Related Work

A similar approach - UA-ZBSA" [10] - tries to classify a set of news headlines into six types of emotions: "anger", "disgust", "fear", "joy", "sadness", "surprise" using the idea that "words which tend to co-occur across many documents with a given emotion are highly probable to express that emotion". From a different point of view, UPAR7 [11] focused on analyzing news articles based on the idea that in a news title all the words can express emotions; therefore, the goal is to identify the main topic within the title.

Furthermore, research was conducted to determine the emotional state of the person who wrote an article. For example, a study was made on blog articles and revealed that angry authors tend to use negative words while joyful ones used more positive words [5]. Other experimental scenarios used blogs where the authors themselves tag an article with their emotional states as baselines in evaluating the system's performance [6].

A more recent approach is SenticNet [12, 13] whose aim is to make conceptual and affective information expressed in natural language more easily accessible even to machines, by enforcing a multi-dimensional hierarchy consisting of a multitude of emotional states with different strengths.

3. Overview of the Architecture

The system's architecture is presented in Fig. 1. The "Current emotional state" module was introduced in the second phase of development and covers the feedback personalization, while all the others tend to focus on identifying the potential emotional state induced by the article, as expressed from the author's point of view.

The system's processing can be divided in 2 steps. Firstly, we perform an objective analysis in the "Core evaluation" module; then, the obtained results are personalized in the "Estimate induced emotional state" module, using input from "Current emotional state" that provides the estimation of the user's state. As it can be observed from Figure 1, the "Core evaluation" module has multiple inputs consisting of the enriched Affective Norms for English Words (ANEW) database [1], a pre-trained Latent Semantic Analysis (LSA) [2] semantic vector space, a corpus of articles that express a dominant emotional state (also called "pure") and the actual article to be evaluated.

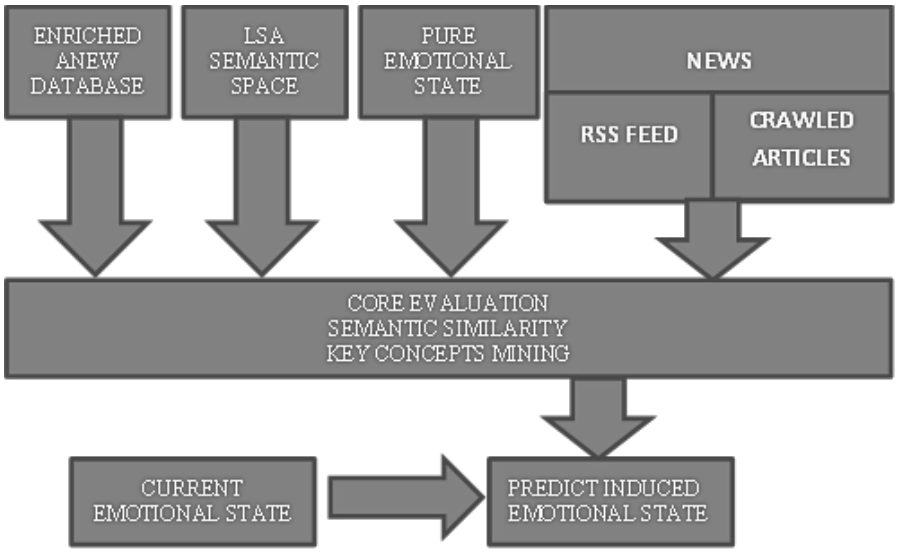


Figure 1. Emo2's Architecture

The ANEW database (containing around 1,000 terms) has a special influence on our system because it gives the scales on which the final results are expressed upon. Furthermore, it contains for each concept 3 dimensions that entangle its potentially induced emotional state (Happy/Unhappy, Calm/ Excited and Controlled/In-Control). The first (affective) and second (arousal) dimensions are considered to form the core emotional state, while the third (dominance) is less strongly related [8]. The Happy/Unhappy axis expresses the presence or absence of a feeling of content, joyfulness. The Calm/Excited dimension shows if a person reading the article is interested in reading more about that news, or might be interested in similar topics. The last dimension shows if the reader feels that he/she is controlled by the information described in the article, or if he/she controls the situation.

The news articles from our experiments were either directly extracted from RSS feeds or crawled from CNN.com using Apache Nutch (http://nutch.apache.org/) with 2 enforced filters. Firstly, a limit of 1 MB was set because many pages had to be ignored due to a large amount of irrelevant content (pictures, links, metadata) and we only extracted the tags corresponding to the title and content. Secondly, a regex was used to limit the crawled domain to general news, extracted from the international version of CNN available at http://edition.cnn.com. In the end, a corpus of around 15,000 articles was built.

Moreover, the objective analysis was further split into two steps: 1/ *semantic similarity* between the contents of articles and the documents used as a baseline (expressing a pure emotional state) by means of LSA and 2/ *mining key concepts*. The "Current emotional state" module uses a questionnaire and a set of 5 similar articles that are evaluated at startup by a reader in order to determine his potential emotional state when reading a news.

4. Context-Independent Evaluation

The context-independent approach contains two dimensions: 1/ an analysis of an article in *comparison* with the *baseline* (documents that express a pure emotional state), therefore a semantic approach, and 2/ a *key concepts mining process* that takes into account the values from the enriched ANEW database, in other words a lexical approach. The documents used as baseline (10 documents per each state) were chosen from general interest news so they could express various emotions. Also, they imply a pure and powerful emotional state. Below is an example of a text expressing a pure state of happiness:

"I visited Egypt last year what an amazing place the food the people seeing the pyramids the mosques the museums it was a magical trip the people there are so friendly it was a tour that I did with other people from all over the world we also did a cruise on the Nile river which was amazing I have been a lot of places but Egypt was one of my favorites what an amazing sight to be in the hotel pool and to see the pyramids in the background".

Semantic similarity between documents is computed using Latent Semantic Analysis (LSA) [2] through the cosine measure between the vector representations of concepts or documents within the semantic space. In order to best fit the specificity of the analysis, the LSA vector space was trained on a large corpus of news articles using the Reuters newswire from 1987, available online at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html. Moreover, LSA has proven to have a similar behavior to human evaluators, so we considered using it to enrich the initial set of concepts from the ANEW database that were experimentally determined through questionnaires. As drawback, LSA is based on a "bag of words" approach [3], meaning that the actual word order or the grammatical structure of a sentence is not taken into account.

In order to address the second approach, focused on key concepts mining, the ANEW database was enriched in 2 steps. Firstly, synonyms for each concept are determined using synsets from WordNet [7], and only the ones stored in the ANEW database are considered to be relevant (if their emotional value is known). Nevertheless, WordNet has a major limitation as it focuses on strict relations between common words having the same corresponding part of speech, so it cannot be used alone as a method to enrich data based on the similarity between concepts. By referring solely to the synonymy relationships from WordNet, the values for a new concept are computed using a weighted mean of the values of its synonyms. Secondly, LSA is also used to determine the closest

concepts from within the enriched database in order to express the valences for a new concept through co-occurrence induced relationships. The actual number of closest concepts from within the pre-trained LSA space considered for expressing the valence of a new word was determined experimentally and, after multiple iterations, we opted to choose the three most dominant and similar concepts.

The overall valences of each sentence are determined as the normalized sum of valences of contained words. Afterwards, the individual results for each article are computed using Equation 1 that gives a greater value to the concepts from the title, rather to the ones from the contents of the article. Since we are dealing with rather short articles whose titles are most often suggestive, we considered that the title is designed to attract the potential users' attention, so it conveys more valuable information than the actual content of the article.

$$value_{key-concepts} = p * TitleValue + (1-p)ContentValue \quad (1)$$

ContentValue and *TitleValue* contain key-concepts valences, with all concepts having the same importance. p was determined experimentally by using increments of 0.05 within the range [0.55; 0.9] and 0.6 proved to be the best alternative for weighting the message in terms of the title, while considering the overall mood induced by the content.

The final results for the context-independent approach are obtained by using Equation 2 that augments the importance of the values obtained through key concepts mining, as the semantic approach is limited by the bag of word approach from LSA:

$$finalResult = p * value_{key-concepts} + (1-p)value_{\ setOfDocuments} \quad (2)$$

The value of p was determined experimentally using values from [0.55; 0.9] and was set to 0.65.

5. Results Personalization

This second phase is used to adapt and personalize the previous results from the context independent evaluation based on the estimation of the current emotional state of the person reading the article. For example, a person who is unhappy tends to be more affected by sad news than someone who is already happy. Also, another factor that proved to be relevant is geo-localization, as people directly involved in the presented situation will be more influenced by it (for example if an accident took place nearby the user's current location). From a more general perspective, the effects of reading a news article are also inter-linked with the reader's psychological traits and further experiments will try to take into consideration more dimensions that impact the estimation of the reader's emotional state.

Taking into account the previous observations, a user must evaluate 5 news articles (see Figure 2) chosen to be as similar as possible in terms of LSA similarity to the news displayed in the end. In other words, for visualizing the main news that will be displayed to the user and their corresponding valence estimations, s/he must beforehand evaluate the top five most similar articles from the corpus.

Additionally, in order to fine-tune even further the displayed results, the user must introduce the continent where he/she lives in. The concepts from the news are analyzed with a geo-localization API (GeoNames - http://www.geonames.org/) that gives access to over 8 million toponyms (various kinds of geographical concepts: countries, cities, continents etc.) and to functions to retrieve the connections between them.

Afterwards, our system determines whether the article refers to a location or event from the continent introduced by the user. In case of a positive result, the values for an article are modified so the ones greater than the median value are augmented, while the others are decreased.

$$NewValue = OldValue\left(1 \pm \frac{Nr.of\ toponyms\ from\ article\ on\ user's\ continent}{Nr.of\ all\ toponyms\ from\ article}\right) \quad (3)$$

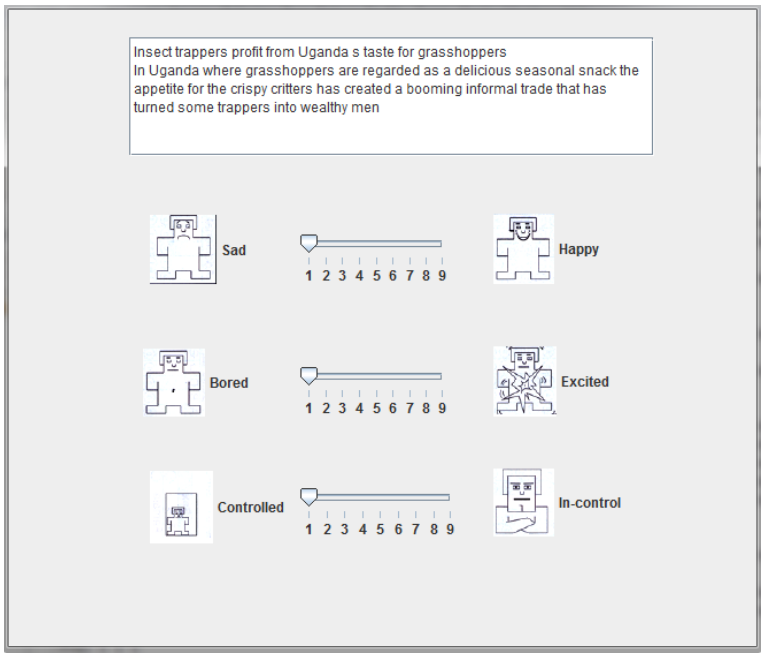
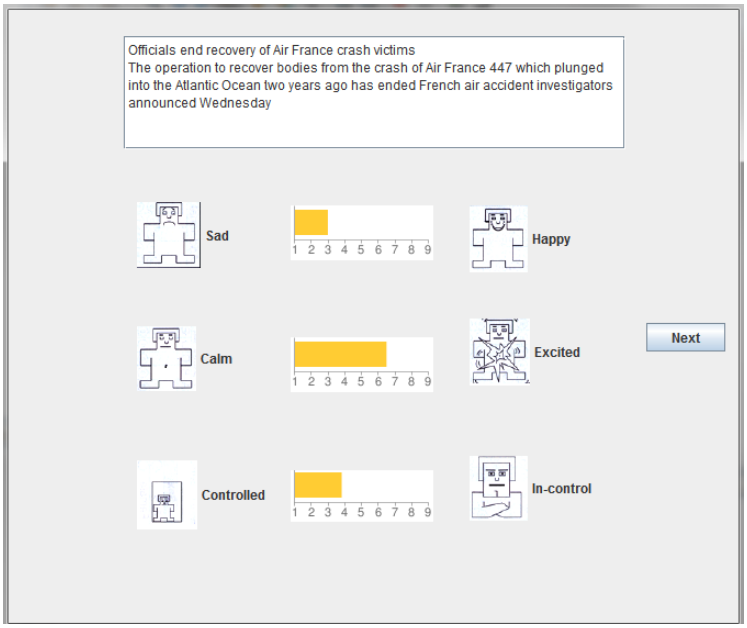


Figure 2. Graphical interface for the initial assessment of articles later used in the personalization step

6. User Interface and Results

An output sample for a given news article is displayed in Figure 3. The 3 dimensions are each represented as a progress bar positioned between the two images of the corresponding pure emotional states.



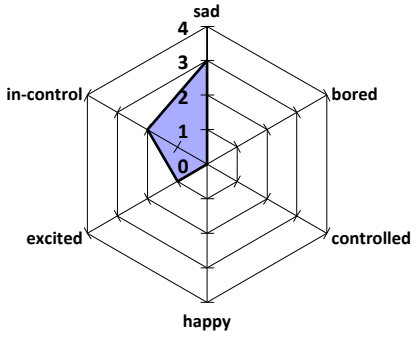


Figure 3. Graphical interface providing feedback on a specific article

We have also opted for presenting the results as a radar graph that introduces the valences for each dimension. Only the dominant subcomponent for each axis is presented on a [0; 4] scale (for example, in Figure 4 we have an article with sad = 3, excited = 1 and in-control = 2).

The results were validated by performing a comparison with the values obtained from a survey conducted on 10 participants from different backgrounds (educational science and computer science). The survey simulated the interaction of a user with the application: firstly, each person was asked to assess 5 news articles that were automatically evaluated to be the closest ones from the

corpus, to the ones that were later on presented to the user; afterwards, the user provided details about his/her geographical location and was asked to manually annotate 10 news articles on the 3 dimensions.

A preliminary drawback of our method was highlighted from the evaluation of the latest article (the 10^{th}) whose semantic similarities with all the documents from our corpus were low (less than 0.1). In this context, this induced noise within our initial assessment, as correlations were rather low (less than 0.2). After eliminating this article from the evaluation based on objective reasons, as it covered different topics from the ones contained within our "pure" emotional documents, the results dramatically improved. Nevertheless this can be considered a positive aspect, as this exception within the evaluation corpus clearly showed means of improvement of the analysis and a clear indicator when the evaluation is prone to noise (the semantic similarity is below a minimum threshold).

While considering only 9 news articles for the context independent evaluation, correlations were computed between the automatic valences determined by our system and the mean valences of all evaluators. The reason why these computations were performed resides in the fact that multiple perspectives from different users converge to the overall induced emotional state. The Pearson correlations are as follows: $r_{affective}$=.44, $r_{arousal}$=.42 and $r_{dominance}$=.47.

For the second phase that presumes the personalization of results, the initial 5 news articles were used to adapt the results for the latter 9 articles. In this case we have opted to perform a correlation for each user between his/her personalized scores and the manual annotations. In order to better grasp the results, the mean and standard deviation for each dimension were determined: $\mu(r_{affective})$=.49, $\sigma(r_{affective})$=.28, $\mu(r_{arousal})$=.16, $\sigma(r_{arousal})$=.17, and $\mu(r_{dominance})$=.36, $\sigma(r_{dominance})$=.31. There are some situations where the second approach has significantly inaccurate results compared to the context-independent approach. A possible explanation is the subjectivity of the evaluators participating at the survey or the misunderstanding of the interpretation of the arousal and dominance dimensions that are harder to grasp in terms of affect estimations. Moreover, by considering the underlying computations, another plausible explanation takes into account that the news from the initial questionnaire were not similar enough to the ones from the results, so the algorithm was misguided and, instead of adjusting the values, it modifies them incorrectly.

7. Conclusions

We have implemented an automatic system that predicts how a person is affected by reading a news article and preliminary validation results, despite the intrinsic subjectivity of the task at hand, are encouraging. We firstly focused on the modules that compute an objective, context-independent value for a news article and then we developed a personalization module that adapts the results according to the emotional state of the person when reading the article. Several natural language techniques, with emphasis on Latent Semantic Analysis, were integrated within our approach. Moreover, based on the conducted survey we have identified exceptions that needed to be addressed and means of improving the overall results.

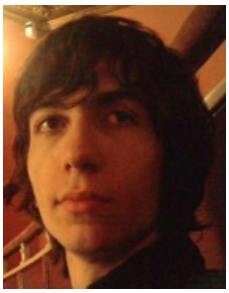
As future research, we intend to conduct more surveys in order to have more feedback and also to add a trust management module to make sure the validation of the system's results is done adequately. For example, if a user rates two similar news very differently (on purpose), the results will not be relevant as the system's logic is based on how similar news are evaluated. Also, not only the results might not be relevant, but also less accurate than without the personalization module. This can create the false impression that the system is faulty and our aim is to limit as much as possible this side effect. Also, we consider that increasing the corpus of "pure" emotional documents can lead to more accurate results, as they are considered a baseline in our analysis.

References

[1] Bradley, M.M., and Lang, P.J. (1999), *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. Gainesville (FL): The Center for Research in Psychophysiology, University of Florida, Tech. Report.

[2] Landauer, T.K., and Dumais, S.T. (1997), *A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge*. In Psychological Review, 104(2), pp. 211–240.
[3] Foltz, P.W. (1996), *Latent semantic analysis for text-based research*. In Behavior Research Methods, Instruments and Computers, 28(2), pp. 197–202.
[4] Kamvar, S.D., Schlosser, M.T., and Garcia-Molina, H. (2003), *The EigenTrust Algorithm for Reputation Management in P2P Networks*. In proceedings of WWW2003, Budapest, Hungary.
[5] Gill, A.J., French, R.M., Gergle, D., and Oberlander, J. (2008), *The Language of Emotion in Short Blog Texts*. In proceedings of CSCW '08, San Diego, pp. 299–302.
[6] Leshed, G., and Kaye, J.J. (2006) *Understanding How Bloggers Feel: Recognizing Affect in Blog Posts*. In proceedings of CHI 2006, Montreal, Canada.
[7] Miller, G.A. (1995), *WordNet: A lexical database*. Communications of the ACM, 38 (11), pp. 39-41.
[8] Mehrabian, A., and Russell, J.A. (1974), *An approach to environmental psychology*. Cambridge, MIT Press.
[9] Lupan, D., Dascalu, M., Trausan-Matu, S., & Dessus, P. (2012). Analyzing emotional states induced by news articles with Latent Semantic Analysis. In A. Ramsay & G. Agre (Eds.), Proc. 15th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2012), LNCS 7557, pp. 59–68.
[10] Kozareva, Z., Navarro, B., Vazquez, S., and Montoyo, A. (2007), *UA-ZBSA: A Headline Emotion Classification through Web Information*. In: Proceeding of the 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 334–337.
[11] Chaumartin, F.-R. (2007), *UPAR7: A knowledge-based system for headline sentiment tagging*. In: Proceeding of the 4th International Workshop on Semantic Evaluations, SemEval 2007, pp. 422–425.
[12] Cambria, E., Grassi, M., Poria, S., and Hussain, A. (2013). *Sentic computing for social media analysis, representation, and retrieval*. In: N. Ramzan et al. (eds.) Social Media Retrieval, ch. 9, pp. 191-215, Springer,
[13] Cambria, E., Song, Y., Wang, H., and Howard, N. (2013). Semantic multi-dimensional scaling for open-domain sentiment analysis. In press: IEEE Intelligent Systems.

Diana LUPAN graduated from University Politehnica of Bucharest, Faculty of Automatic Control and Computers in 2011 and is currently in the last year of her master degree in Internet Systems Engineering. Her research interests are centered on applying specific natural language processing techniques for extracting emotions from written texts. She has already published in this research area in multiple national and international conferences (AIMSA and RoCHI).



Stefan BOBOCESCU-KESIKIS has graduated University Politehnica of Bucharest in 2011 and is studying for his master degree in Internet System Engineering. He is passionate about distributed computing, information retrieval and affective computing.

Mihai DASCALU was head of promotion 2009, University Politehnica of Bucharest, and currently holds 2 master degrees - one in Internet systems engineering, UPB, and one in knowledge extraction, University of Nantes. He's in the last year of his double PhD in computer science (UPB) and educational sciences (University Grenoble Alpes) and has experience in national and international projects (FP7 LTfLL, FP7 ERRIC and CNCSIS K-TEAMS). He has more than 40 published papers, including top computer education conferences (ITS, CSCL) and other renowned international conferences (ICALT, EC-TEL, ICWL, ISPDC, AIMSA). Complementary to his competencies in natural language processing, technology-enhanced learning and discourse analysis, Mihai holds a multitude of professional certifications (e.g. PMP, RMP- PMI, CEH, CISSP).

Stefan Trausan-Matu, PhD (http://www.racai.ro/~trausan) is a full professor at the Computer Science Department of the "Politehnica" University of Bucharest, and principal researcher at the Institute of Artificial Intelligence of the Romanian Academy. He is lecturing Algorithms Analysis and Design, Human-Computer Interaction, Natural Language Processing, Adaptive and Collaborative Systems. He was a Fulbright post-doc at Drexel University, Philadelphia, USA, was invited professor and lectured in USA, Netherlands, France, San Marino, Germany, Puerto Rico, etc. His current research interests are: Discourse Analysis, Intertextuality, Creativity Fostering, Computer-Supported Collaborative Learning, Human-Computer Interaction, and Philosophy. Prof Trausan-Matu has authored or edited 17 books, 25 book chapters and more than 200 peer-reviewed papers.

Philippe DESSUS, PhD, is a full professor in educational sciences at the Teacher Education Institute (IUFM) of Univ. Grenoble Alpes. He is head of the Laboratory of Educational Sciences in the same university. He was a former post-doc researcher in educational sciences at Liège University (Belgium), as well as an invited researcher at TECFA, Geneva University (Switzerland) and at LIG-MeTAH, Univ. Grenoble Alpes. Philippe Dessus was involved in the LTfLL EC project (7th PCRD-STREP) and in several French research projects. His research interests are to design and implement experiments involving ICTs in educational contexts, related to two main themes: computer-aided reading/writing environments and instructional design. Most of these experiments have used Latent Semantic Analysis as a tool for delivering automated assessments to the learners. He has published over 40 international articles in conferences, proceedings, invited talks or reports on that theme.