LREC 2016 Workshop

Emotion and Sentiment Analysis
PROCEEDINGS

Edited by

J. Fernando Sánchez-Rada and Björn Schuller

23 May 2016

Proceedings of the LREC 2016 Workshop
"Emotion and Sentiment Analysis"

23 May 2016 – Portorož, Slovenia

Edited by J. Fernando Sánchez-Rada and Björn Schuller

Acknowledgments: This work has received funding from the EU's Horizon 2020 research and innovation programme through project MixedEmotions (H2020 RIA grant agreement #644632).

Organising Committee

J. Fernando Sánchez-Rada*	UPM, Spain
Björn Schuller *	Imperial College London, United Kingdom
Gabriela Vulcu	Insight Centre for Data Analytics, NUIG, Ireland
Carlos A. Iglesias	UPM, Spain
Paul Buitelaar	Insight Centre for Data Analytics, NUIG, Ireland
Laurence Devillers	LIMSI, France

*: Main editors and chairs of the Organising Committee

Programme Committee

Elisabeth André	University of Augsburg, Germany
Noam Amir	Tel-Aviv University, Israel
Rodrigo Agerri	EHU, Spain
Cristina Bosco	University of Torino, Italy
Felix Burkhardt	Deutsche Telekom, Germany
Antonio Camurri	University of Genova, Italy
Montse Cuadros	VicomTech, Spain
Julien Epps	NICTA, Australia
Francesca Frontini	CNR, Italy
Diana Maynard	University of Sheffield, United Kingdom
Sapna Negi	Insight Centre for Data Analytics, NUIG, Ireland
Viviana Patti	University of Torino, Italy
Albert Salah	Boğaziçi University, Turkey
Jianhua Tao	CAS, P.R. China
Michel Valstar	University of Nottingham, United Kingdom
Benjamin Weiss	Technische Universität Berlin, Germany
Ian Wood	Insight Centre for Data Analytics, NUIG, Ireland

Preface

ESA 2016 is the sixth edition of the highly successful series of Corpora for Research on Emotion. As its predecessors, the aim of this workshop is to connect the related fields around sentiment, emotion and social signals, exploring the state of the art in applications and resources. All this, with a special interest on multidisciplinarity, multilingualism and multimodality. This workshop is a much needed effort to fight the scarcity of quality annotated resources for emotion and sentiment research, especially for different modalities and languages.

This year's edition once again puts an emphasis on common models and formats, as a standardization process would foster the creation of interoperable resources. In particular, researchers have been encouraged to share their experience with Linked Data representation of emotions and sentiment, or any other application of Linked Data in the field, such as enriching existing data or publishing corpora and lexica in the Linked Open Data cloud.

Approaches on semi-automated and collaborative labeling of large data archives are also of interest, such as by efficient combinations of active learning and crowdsourcing, in particular for combined annotations of emotion, sentiment, and social signals. Multi- and cross-corpus studies (transfer learning, standardisation, corpus quality assessment, etc.) are further highly relevant, given their importance in order to test the generalisation power of models.

The workshop is supported by the Linked Data Models for Emotion and Sentiment Analysis W3C Community Group [1], the Association for the Advancement of Affective Computing [2] and the SSPNet [3] – some of the members of the organizing committee of the present workshop are executive members of these bodies.

As organising committee of this workshop, we would like to thank the organisers of LREC 2016 for their tireless efforts and for accepting ESA as a satellite workshop. We also thank every single member of the programme committee for their support since the announcement of the workshop, and their hard work with the reviews and feedback. Last, but not least, we are thankful to the community for the overwhelming interest and number of high-quality submissions. This is yet another proof that the emotion and sentiment analysis community is thriving. Unfortunately, not all submitted works could be represented in the workshop.

J.F. Sánchez-Rada, B. Schuller, G. Vulcu, C. A. Iglesias, P. Buitelaar, L. Devillers May 2016

[1] http://www.w3.org/community/sentiment/
[2] http://emotion-research.net/
[3] http://sspnet.eu/

Programme

9:00 – 9.10	Introduction by Workshop Chair
9.10 – 10:30	**Social Media**
Cristina Bosco et al.	Tweeting in the Debate about Catalan Elections
Ian D. Wood and Sebastian Ruder	Emoji as Emotion Tags for Tweets
Antoni Sobkowicz and Wojciech Stokowiec	Steam Review Dataset - new, large scale sentiment dataset
10:30 – 11:00	Coffee break
11:00 – 13:00	**Corpora and Data Collection**
Ebuka Ibeke et al.	A Curated Corpus for Sentiment-Topic Analysis
Jasy Liew Suet Yan and Howard R. Turtle	EmoCues-28: Extracting Words from Emotion Cues for a Fine-grained Emotion Lexicon
Lea Canales et al.	A Bootstrapping Technique to Annotate Emotional Corpora Automatically
Francis Bond et al.	A Multilingual Sentiment Corpus for Chinese, English and Japanese
13:00 – 14:00	Lunch break
14:00 – 15:00	**Personality and User Modelling**
Shivani Poddar et al.	PACMAN: Psycho and Computational Framework of an Individual (Man)
Veronika Vincze1, Klára Hegedűs, Gábor Berend and Richárd Farkas	Telltale Trips: Personality Traits in Travel Blogs
15:00 – 16:00	**Linked Data and Semantics**
Minsu Ko	Semantic Classification and Weight Matrices Derived from the Creation of Emotional Word Dictionary for Semantic Computing
J. Fernando Sánchez-Rada et al.	Towards a Common Linked Data Model for Sentiment and Emotion Analysis
16:00 – 16:30	Coffee break
16:30 – 18:00	**Beyond Text Analysis**
Bin Dong, Zixing Zhang and Björn Schuller	Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition
Rebekah Wegener, Christian Kohlschein, Sabina Jeschke and Björn Schuller	Automatic Detection of Textual Triggers of Reader Emotion in Short Stories
Andrew Moore, Paul Rayson and Steven Young	Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries

Table of Contents

Regular Papers

Semantic Classification and Weight Matrices Derived from the Creation of Emotional Word Dictionary for Semantic Computing
Minsu Ko 1

PACMAN: Psycho and Computational Framework of an Individual (Man)
Shivani Poddar, Sindhu Kiranmai Ernala and Navjyoti Singh 10

Telltale Trips: Personality Traits in Travel Blogs
Veronika Vincze1, Klára Hegedűs, Gábor Berend and Richárd Farkas 18

A Bootstrapping Technique to Annotate Emotional Corpora Automatically
Lea Canales, Carlo Strapparava, Ester Boldrini and Patricio Martínez-Barco 25

A Curated Corpus for Sentiment-Topic Analysis
Ebuka Ibeke, Chenghua Lin, Chris Coe, Adam Wyner, Dong Liu, Mohamad Hardyman Barawi and Noor Fazilla Abd Yusof 32

EmoCues-28: Extracting Words from Emotion Cues for a Fine-grained Emotion Lexicon
Jasy Liew Suet Yan and Howard R. Turtle 40

Towards a Common Linked Data Model for Sentiment and Emotion Analysis
J. Fernando Sánchez-Rada, Björn Schuller, Viviana Patti, Paul Buitelaar, Gabriela Vulcu, Felix Burkhardt, Chloé Clavel, Michael Petychakis and Carlos A. Iglesias 48

Short Papers

Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries
Andrew Moore, Paul Rayson and Steven Young 63

Tweeting in the Debate about Catalan Elections
Cristina Bosco, Mirko Lai, Viviana Patti, Francisco M. Rangel Pardo and Paolo Rosso 67

Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition
Bin Dong, Zixing Zhang and Björn Schuller 71

Emoji as Emotion Tags for Tweets
Ian D. Wood and Sebastian Ruder 76

Automatic Detection of Textual Triggers of Reader Emotion in Short Stories
Rebekah Wegener, Christian Kohlschein, Sabina Jeschke and Björn Schuller 80

Steam Review Dataset - new, large scale sentiment dataset
Antoni Sobkowicz and Wojciech Stokowiec 55

A Multilingual Sentiment Corpus for Chinese, English and Japanese
Francis Bond, Tomoko Ohkuma, Luís Morgado da Costa, Yasuhide Miura, Rachel Chen, Takayuki Kuribayashi and Wenjie Wang 59

Telltale Trips: Personality Traits in Travel Blogs

Veronika Vincze[1], **Klára Hegedűs**[2], **Gábor Berend**[3], **Richárd Farkas**[3]
[1]MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
[2]Department of Psychology, University of Szeged
klarahegedus92@gmail.com
[3]Institute of Informatics, University of Szeged
{berendg,rfarkas}@inf.u-szeged.hu

Abstract

Here we present a corpus that contains blog texts about traveling. The main focus of our research is the personality trait of the person hence we do not just annotate opinions in the classical sense but we also mark those phrases that refer to the personality type of the author. We illustrate the annotation principles with several examples and we calculate inter-annotator agreement rates. In the long run, our main goal is to employ personality data in a real-world application, e.g. a recommendation system.

Keywords: psycholinguistics, corpus, opinion mining

1. Introduction

Allport (1961) describes personality as "the dynamic organization within the individual of those psychophysical systems that determine his characteristic behavior and thought". According to this definition, in personal psychology, it is well-known that someone's personality may manifest in several ways, e.g. the way he behaves in certain situations, his communication style or his storytelling. Thus, texts authored by the same person include some stylistic or linguistic features that are connected to the author's personality and the linguistic analysis of such texts may reveal what personality type the author belongs to.

Nowadays, the role of social media is becoming more and more significant, especially due to its importance in modern communication. The billions of tweets, wall posts and likes reveal a lot of user preferences, for instance, what type of products they choose, what type of music, books, cars or food they prefer, what destinations they travel to for holiday, what political parties they vote for and so on. All these pieces of data can be exploited in several fields of natural language processing, for instance, in personalized recommendation systems.

In this paper, we present our SzegedTrip corpus of travel blogs written in English, which contains manual annotation for opinions, besides, linguistic markers of the author's personality are also annotated. The author's personality and his/her opinions may correlate: for instance, his/her preference for a specific hotel in a quiet village may be related to his introvert personality and also, the owner of the hotel may identify what type of personality their guests have and hotel's facilities can be improved accordingly etc. In this way, the corpus can be exploited in both computational psychology and opinion mining: the corpus makes it possible to experiment with machine learning tools to identify the textual markers of personality and opinionated phrases and later on, the detection of what personality type the author may have. On the other hand, recommendation systems may also profit from the corpus.

2. Related Work

Here we summarize the most important studies on sentiment analysis and opinion mining, as well as personal psychology related to travel personality.

2.1. Sentiment analysis and opinion mining

Sentiment analysis and opinion mining aim at making inferences about someone's feelings (towards a given subject, e.g. a trip).

From a travel-related point of view, some authors (Ye et al., 2009; Ye et al., 2011) used opinion mining techniques to test the impact of consumer-generated travel reviews on hotel bookings. On the other hand, it can be useful for travel agents to identify someone's travel personality. With this information it would be possible to make preferable destination recommendations, so the advertising policy could be more targeted and personalized. Our corpus makes it possible to investigate the relationship of textual markers and the author's personality. In both cases, the main goal is to collect information from textual clues about the belief and behavioral patterns of the person in question.

There are some annotated corpora for sentiment analysis and opinion mining:

- The MPQA corpus contains newswire texts and annotates sources (the holder of the opinion), targets of the opinion and subjectivity (Wilson and Wiebe, 2005).
- The J. D. Power & Associates corpus (Kessler et al., 2010) contains automotive review blog posts, where named entities are annotated for sentiment towards them. Linguistic modifiers and markers of polarity are also annotated. Sources that do not coincide with the author of the text are also separately marked.
- Sayeed et al. (2011) present a corpus of information technology articles, which are annotated for linguistic markers of opinions at the word level.
- Scheible and Schütze (2013) distinguish between subjectivity and sentiment relevance. They label sen-

tences as sentiment relevant if it contains some information on the sentiment that the document conveys.

2.2. Travel personality

Our personality contains many permanent traits which predict our behavior in many situation. Personality impacts brand preference, product choice and also travel-related decisions too (Yoo and Gretzel, 2011; Cao and Mokhtarian, 2005), for example the choice of destination and the organization of programs, activities during the holiday (Yoo and Gretzel, 2011).

In personal psychology, the Five Factor Model of Personality is one of the most common personality theories. According to the Big Five Model (see Mccrae and Costa (1987)), there are five determinative personality dimensions. These are: openness, conscientiousness, extraversion, agreeableness, and neuroticism. One trait indicates a spectrum, so there are high and low levels of these dimensions. In the case of travel-related reviews, some of these traits could be easily identified. For example, an individual with high level of openness would try to learn as much as possible about the local culture; and a conscientious person would plan every detail of the trip in advance.

However, it is important to take into consideration that the tendency of writing an online review is also related to personality traits. Agreeableness, openness, conscientiousness and/or extraversion are related to knowledge sharing intentions, while neurotic individuals would less likely be involved in consumer-generated media.

Besides the Big Five Model, there are some models especially formed for travel personalities. For example, Pearce's (1988) "travel career ladder" refers to tourist motivation as a changeable state, based on Maslow's hierarchy of needs; Cohen's (1972) "strangeness-familiarity" model takes place in a broader, social context; Salomon and Mokhtarian's model (1998), which suggest a number of reasons why people travel and Plog's "travel personality" model.

Plog's model (2001) analyzes in detail the relationship of personality traits and traveling habits. The model contains five types through a spectrum: venturer, near venturer, midcentric, near dependable and dependable. He describes a dependable individual as a cautious, conservative, intellectually restricted person, who prefers popular, well-known products, could not make his/her own decisions, faces daily life with low activity level, likes structure, likes to be with his/her family and friends. As for a dependable's travel habits, he/she travels less frequently for shorter periods of time, prefers to stay in cheaper hostels and motels with his/her relatives, selects recreational, relaxing activities, selects well-defined, escorted tours, likes touristy spots, returns to well-tried destinations again and again. In contrast, a venturer person is curious, energetic and active, makes decision quickly, likes to choose new products, fills the trip with varying activities and challenges. A venturer travels more frequently for longer periods of time, prefers unusual destinations and unconventional accommodations, prefers to participate in local customs and habits and organizes exciting activities.

2.3. Identifying personality

Recently, there has been a shared task aiming at computational personality recognition (Celli et al., 2013). They released two datasets – essays and a subset of the myPersonality dataset –, which include gold standard personality labels and texts (essays and Facebook status updates) written by the persons themselves.

Yerva et al. (2013) present their recommendation system for landmarks at a given place, based on global and user-specific ranking model. They make use of the user's likes and posts and friends' activities on Facebook.

The main contributions of our new corpus are the following. It contains blog texts about traveling, which is – to the best of our knowledge – a new domain in sentiment analysis. Although there has been some previous work on opinion mining related to traveling, e.g. Ye et al. (2011) and Kasper and Vela (2011) annotated travel related opinions, (e.g. the target, the polarity, the aspect, the holder and the time of the opinion), the main focus of our research is not just the person's opinion towards a given subject but the personality trait of the person as well. Similar to Scheible and Schütze (2013) but in contrast with MPQA (Wilson and Wiebe, 2005) and JDPA (Kessler et al., 2010), we do not just annotate opinions in the classical sense, i.e. expressing certain views about some targets: we also mark those phrases that refer to the personality type of the author. In the long run, our main goal is to employ personality data in a real-world application, e.g. a recommendation system, where we aim at exploiting the psychological profile of the user when proposing travel destinations to him.

3. The Corpus

We collected 500 blog entries which describe trips made by their authors. It was important to access more than one post from one author, so instead of collecting from global travel review databases (like Ye et al. (2009) and Nakayama and Fujii (2013)), we had to use personal blogs. Like Ye et al. (2009), we pre-established some popular areas, so we collected reviews related to them. Trips targeted one of the five following destinations: Barcelona, Hungary, India, Los Angeles and Middle East countries.

Blog entries were collected with the help of queries including words related to travelling and one of the destinations like "trip to Hungary", "journey in China" etc. However, a lot of data collected in this way turned out to be unrelated to travelling, so later on, we manually filtered those blogs that had nothing to do with travelling.

There are 100 blog entries belonging to each destination in the corpus. Besides, we also collected other types of texts which were authored by the same people since we believe that they can also be exploited in identifying the personality type of the author and later on, we would like to annotate them as well for linguistic markers of personality traits.

4. Annotation Principles

The SzegedTrip corpus was manually annotated by a student of psychology, who was instructed to mark sentences or clauses which contain information useful for determining the author's (travelling) personality. These may be sentences that express the author's positive or negative opinion

on a certain target (which is present in the sentence) and targetless sentences as well. In the latter case, it is rather the whole situation or event that invokes some feelings rather than a specific thing/person/entity. Factual sentences may be also included even if they do not contain polar / subjective terms but they are relevant and suggestive of a positive or negative opinion.
It is primarily the relevance of content that counts when selecting the sentence for annotation (rather than the exact wording, the presence of polar or subjective terms, the usage of certain syntactic structures etc.). Opinions can be understood in this way (similar to Sayeed et al. (2011)):

> A expresses an opinion (about B) if an interested party C may be affected by A's words.

In the traveling context, B is the target of opinion, e.g. a hotel, a city, a restaurant, a meal etc. B may not always be present in the sentence /clause as in:

> *My hotel room was small but had a wonderful view on the sea.*

Here the first clause contains a negative opinion on the hotel room and the second clause contains a positive opinion on the same target, however, at the second time it is not repeated.
We employed hierarchical (two-level) annotation. At first, we annotated three kinds of opinions (first-level annotation):
Targeted positive opinions:

> *We visited the Place des Vosges, which is now a very nice park.*
>
> *Experience Music Project - thank you, Paul Allen. This is a shrine to music in a gorgeous Frank Ghery-designed building.*

Targeted negative opinions:

> *The portions were on the small side.*
>
> *The morning greeted us with heavy rain clouds and a big dip in temperature.*

Targetless opinions:

> *Unfortunately you can not be on the top deck during this cruise or you may meet the guillotine.* (Here, the author does not like the restrictions on being on the top deck although he might still like the cruise.)
>
> *My reservation had been canceled due to something wrong with my credit card when I bought the ticket.* (The problem is with the airline or its reservation system, which the author does not like.)
>
> *My luggage was lost on the flight.* (The problem is with the airline losing some luggage.)

At the second level of annotation, we annotated the target of the opinion and phrases that are linguistic markers of the given opinion (descriptor). Each opinion should have exactly one target and at least one descriptor (with some exceptions). As more than one opinion may belong to a specific target, moreover, they can be situated in the text far away from each other, targets referring to the same entity are marked with the same number (similar to coreference annotation). Below, only second-level annotation is marked: targets are bold and descriptors are underlined.

> ***Hot food*** *consisted on scrambled eggs, which were cooked to my taste, bacon, which was very tasty, but fattier that I like, stewed tomatoes that were very good, boiled rice and chicken soup.*
>
> ***Experience Music Project*** *– thank you, Paul Allen. This is a shrine to music in a gorgeous Frank Ghery-designed building.*
>
> *The* ***portions*** *were on the small side.*

In some cases, we mark the target more than once in the sentence because the first mention of the target is objective and the part of the sentence which includes the opinion uses only a pronominal reference to the target as in:

> *The* ***hotel*** *was located downtown and* ***it*** *was one of the worst I've ever seen.*

We mark textual parts as personality markers which are not direct opinions but are related to the author's "travel personality". When collecting the important details of travel personality, we take into consideration Plog's model (2001) and partly the Big Five dimensions (Mccrae and Costa, 1987).
According to Plog's model (2001), these phrases may be useful in e.g. figuring out whether the author:

- Likes traveling alone or with others;
- Likes organizing his/her own trip;
- Likes traveling with a traveling agency;
- Likes stability and well-known sites (similar to home);
- Likes long journeys (in time and in place as well);
- Is a frequent traveler;
- Likes going around during his/her holiday;
- Likes staying at a fixed place during his/her holiday;
- Prefers big cities, countryside, seaside, exotic places...
- Prefers flying, traveling by car or by train...

On the other hand, we also annotate expressions as personality markers which are related to the Big Five dimensions of personality (see Mccrae and Costa (1987)). Some examples for personality markers:

I uploaded a few facebook photos. (The author likes informing others, which indicates extraversion.)

In the tourist room Americans were far outnumbered by Japanese and Arabs/Moslems. (The author does not like unfamiliar situations, so he may not be open to new things or experiences.)

For each personality marker, we also annotated it according to Plog's model (i.e. it refers to a venturer or a dependable) or the Big Five model (i.e. it encodes openness, extraversion, agreeableness, conscientiousness or neuroticism). It might also occur that the very same blog text contains different dimensions of the same personality marker, which indicates that people's personality cannot be described with a one-dimensional approach: rather, it is also essential what aspect is connected to the given personality marker (e.g. someone likes to taste new meals, which refers to his openness from a gastronomic point of view but he usually spends his holiday in the same hotel, which reflects his conservativeness concerning accommodation). This fact also demonstrates that aspect-oriented opinion mining might be successfully exploited in computational psychology.

5. Statistical Data on the Corpus

The corpus contains 500 blog entries, approximately 20,000 sentences and 400,000 tokens. Basic statistical data on the frequency of each annotated category can be seen in Table 1. Concerning opinions, it is revealed that people mostly express their positive opinions in their blogs, that is, they prefer writing about what they liked. This is highlighted by the percentage rates of positive and negative opinions and descriptors as well: at least 83% of the opinions and descriptors are positive. There is only one exception to this tendency: blogs about journeys to India tend to contain more negative opinions, which may be due to the fact that India is very dissimilar to Western countries and people tend to cope with the gaps between their home culture and that of India to a lesser degree than at the other destinations.

In the blogs, there are 4315 targets mentioned in 4481 opinions, which means that some opinions do not include an explicit linguistic marker for the target (for instance, if it coincides with the subject of the previous clause, the subject may be omitted in elliptic sentences). However, each opinion contains 1.42 descriptors on average, which suggests that people usually express their views with more than one descriptor, most probably, they want to emphasize their likes or dislikes in this way.

As for personality markers, texts were annotated with the Big Five categories and/or Plog's categories. Table 2 shows the results. For the Big Five categories, we also made a distinction between higher and lower levels of each dimension: the number of occurrences denoting the high dimension of each trait is marked at the left hand side of the slash and the low dimension at the right hand side.

The data in Table 2 reveal some interesting tendencies. For instance, we can find more manifestations of a dependable personality than those of venturers: about 38% of the markers refer to a venturer. However, there are notable differences for the destinations (results are significant: χ^2-test, p = 0.0073): for instance, the rate of dependables and venturers is about 50-50% in the case of Hungary, so according to our dataset, the most probable destination a venturer has chosen is Hungary.

As for the Big Five categories, we can again find some significant differences among the destinations (χ^2-test, p = 0.0003). For instance, it is mostly travelers to Barcelona that express their extraversion in their blogs and agreeableness can be typically discovered in texts about India. In general, most of the markers are related to extraversion but neuroticism does not seem to be a frequent category, hence it may be concluded that travel blogs are not indicative of the person's neuroticism level but they can be suggestive of the person's extraversion level.

	A1	A2	A3	A4	A5
A1		31.09	26.97	19.50	30.44
A2	31.09		21.81	16.20	31.52
A3	26.97	21.81		19.29	37.42
A4	19.50	16.20	19.29		21.85
A5	30.44	31.52	37.42	21.85	

Table 3: Agreement rates in terms of micro F-scores.

6. Inter-annotator Agreement Rates

In order to test the difficulty of the task and to calculate inter-annotator agreement rates, 10 texts from each destination were annotated by four more annotators. All of them were trained linguists and could speak English at a high level. Annotators worked on texts independently and if in need, they could turn to the annotation guidelines summarized in Section 4., besides, they could consult with the chief annotator who was responsible for creating the guidelines and for supervising the annotation work process.

For calculating pairwise inter-annotator agreement rates, the metric F-score was used. We applied a very strict evaluation methodology here: we accepted an annotated phrase as true positive if and only if the same snippet of text was marked by both annotators (with exact boundary matches) and it was labeled in the same way. For instance, if one annotator marked the phrase *it took long to get coffee* and the other one marked *took long to get coffee* (i.e. without marking "it"), it counted as an error in the evaluation. In other cases, the lack of marking a conjunction led to annotation mismatches as in *(and) we docked in Rhodes instead, which I might add was very lovely*.

Aggregated inter-annotator agreement rates can be seen in Table 3 in terms of micro F-scores, and agreement rates calculated for each category separately are shown in Tables 4 to 7.

Based on the agreement rates, it is revealed that while four annotators could achieve approximately the same level of agreement in each scenario, the fifth one was somewhat behind them and obtained lower scores. This might be related to the fact that she had the least experience with annotating English texts, which might have influenced her

	Barcelona	**Hungary**	**India**	**Los Angeles**	**Middle East**	**Total**
Op	988	831	850	930	882	4,481
PosOp	821 (83.10)	706 (84.96)	632 (74.35)	801 (86.13)	738 (83.67)	3698 (82.53)
NegOp	167 (16.90)	125 (15.04)	218 (25.65)	129 (13.87)	144 (16.33)	783 (17.47)
Desc	1,478	1,148	1,256	1,226	1,251	6,359
PosDesc	1,241 (83.96)	965 (84.06)	921 (73.33)	1,064 (86.79)	1,047 (83.69)	5,238 (82.37)
NegDesc	237 (16.04)	183 (15.94)	335 (26.67)	162 (13.21)	204 (16.31)	1,121 (17.63)
Target	947	806	829	892	841	4,315
PersMark	358	250	308	235	315	1,466
Sentence	4,152	3,644	3,769	4,170	3,926	19,661
Token	87,624	79,386	76,533	83,161	83,266	409,970

Table 1: Statistical data on the annotated categories. Op: opinion, Desc: descriptor, Pos: positive, Neg: negative, PersMark: personality marker.

	Barcelona	**Hungary**	**India**	**Los Angeles**	**Middle East**	**Total**
Venturer	80	88	75	44	68	355
Dependable	149	95	104	96	133	577
Extraversion	55/0	17/1	15/7	20/2	26/3	133/13
Agreeableness	3/0	3/0	11/0	3/0	3/0	23/0
Openness	10/0	15/0	16/0	8/0	23/0	72/0
Conscientiousness	10/5	5/4	8/4	2/2	4/3	29/18
Neuroticism	1/0	0/0	0/0	0/0	2/1	3/1

Table 2: Statistical data on personality markers (high dimension/low dimension of the trait).

Pos	A1	A2	A3	A4	A5
A1		25.50	26.39	15.69	29.73
A2	25.50		23.54	18.27	29.04
A3	26.39	23.54		20.36	43.54
A4	15.69	18.27	20.36		25.95
A5	29.73	29.04	43.54	25.95	

Neg	A1	A2	A3	A4	A5
A1		22.83	23.00	11.43	27.32
A2	22.83		14.70	3.60	24.11
A3	23.00	14.70		18.90	48.41
A4	11.43	3.60	18.90		15.15
A5	27.32	24.11	48.41	15.15	

Table 4: Agreement rates for positive and negative opinions.

	A1	A2	A3	A4	A5
A1		37.38	28.23	24.48	28.87
A2	37.38		27.53	22.22	35.00
A3	28.23	27.53		18.94	45.64
A4	24.48	22.22	18.94		18.44
A5	28.87	35.00	45.64	18.44	

Table 5: Agreement rates for targets.

work.

It is revealed from the results that annotators can achieve a higher agreement rate in the case of opinions than in the case of personality markers, which might imply that the latter is even more subjective. However, the difference is not tremendous and thus, the difficulty of annotating personality markers is comparable to other semantics-related tasks like marking of opinions.

Pos	A1	A2	A3	A4	A5
A1		31.70	26.32	16.80	30.20
A2	31.70		17.27	13.81	33.25
A3	26.32	17.27		16.56	24.43
A4	16.80	13.81	16.56		19.19
A5	30.20	33.25	24.43	19.19	

Neg	A1	A2	A3	A4	A5
A1		18.87	16.27	9.76	21.40
A2	18.87		12.75	11.32	26.90
A3	16.27	12.75		8.09	11.89
A4	9.76	11.32	8.09		12.87
A5	21.40	26.90	11.89	12.87	

Table 6: Agreement rates for positive and negative descriptors.

Based on the results, we conclude that the strict evaluation methodology might be one reason for the modest agreement rates. In order to test this hypothesis empirically, we manually evaluated the positive opinions marked by those annotators who could reach the highest inter-annotator agreement rate (i.e. A3 and A5 with an F-score of 43.54). Throughout the manual evaluation, we accepted as true positives the cases similar to the above mentioned examples. For instance, it was typical that one of the annotators marked some text spans as one opinion while the other one separated them into two opinions: the phrase *Dinner was excellent with a delicious pork dish on the menu* was marked as one opinion by one annotator but the other

	A1	A2	A3	A4	A5
A1		3.24	9.92	13.45	19.73
A2	3.24		1.79	3.65	4.90
A3	9.92	1.79		11.60	28.35
A4	13.45	3.65	11.60		12.45
A5	19.73	4.90	28.35	12.45	

Table 7: Agreement rates for personality markers.

one split it into two, marking one opinion on the dinner as a whole meal and another one on the pork dish, which both can be acceptable solutions.

With this lenient evaluation methodology, the agreement rate we obtained was 76.52 in terms of F-score, which is on a par with the sentence-level agreement rates reported for the MPQA corpus (Wiebe and Cardie, 2005). Hence, we believe that strict boundary matches may be refined and some more relaxed methodology should be applied to the automatic evaluation of such semantics-related tasks. For instance, only the head of the target phrase should be matched and the exact boundaries of the annotated phrases do not need to be the same.

7. Possible Uses of the Corpus

First of all, our corpus can be used as training and evaluation database for machine learning algorithms that are designed to detect personality traits and opinions. As the inter-annotator agreement rates indicate, marking opinions and personality markers is a subjective task by its nature, similar to other semantics-related NLP tasks (e.g. machine translation or information retrieval) where there are multiple solutions that might be acceptable. In such cases, multiple good solutions are taken into account when evaluating the performance of an automatic system. For instance, the scores BLEU and ROUGE are computed on the basis of comparing the system's output to multiple human solutions (Papineni et al., 2002; Lin, 2004) and the union and intersection of keyphrases given by different annotators are used as gold standard in opinionated keyphrase extraction (Berend and Vincze, 2012). In harmony with these evaluation methodologies, the five different annotations available for a part of our corpus also makes it possible to evaluate automatic methods aiming at detecting personality traits in a more sophisticated way.

Besides, the corpus may be also of use for real-world users. For instance, travellers who aim to travel to one of the destinations described in the corpus can have access to an annotated collection of blog descriptions about the destination they are interested in. Travel agencies may also profit from the corpus. Finally, corpus data may serve as feedback to the owners or workers in hotels and restaurants or those working in tourism at the given place. It can be easily collected from the corpus what those aspects (targets) are that are liked/disliked by most people, which later may determine priorities in development or marketing strategies. To take an example, we carried out a qualitative analysis of targets of negative opinions, which revealed some local specialties. In India, people were mostly dissatisfied with the traffic and dirt, however, in Los Angeles, some reasons for being discontent were that the traveller could not see any celebrities or s/he was annoyed by autograph hunters and in a Middle Eastern country, the traveller did not like that the country was becoming too similar to Western countries and thus losing to some extent its traditional culture. All these differences may be exploited in personalized travel offers, created by either travel agents or automatic systems.

8. Conclusions

In this paper, we presented the SzegedTrip corpus of travel blogs annotated for opinions and linguistic markers of personality. We illustrated the main annotation principles with several examples and we showed that the difficulty of the two tasks is similar, as far as the inter-annotator agreement rates are concerned. However, our experiments also demonstrate that a more relaxed metrics for measuring agreement rates is desirable as opposed to strict boundary matching because of the highly semantic nature of the task. Corpus data can be exploited in personalized offers, either created by human experts or automatic recommendation systems. Besides, the annotated corpus makes it possible to experiment with the automatic identification of the author's personality type, which we would like to implement in the future. The corpus can be freely downloaded from our website (`http://rgai.inf.u-szeged.hu/szegedtrip`).

9. Acknowledgments

Richárd Farkas was funded by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

10. Bibliographical References

Allport, G. (1961). *Pattern and Growth in Personality*. Rinehart & Winston.

Berend, G. and Vincze, V. (2012). How to evaluate opinionated keyphrase extraction? In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 99–103, Jeju, Korea, July. Association for Computational Linguistics.

Cao, X. and Mokhtarian, P. L. (2005). How do individuals adapt their personal travel? Objective and subjective influences on the consideration of travel-related strategies for San Francisco Bay Area commuters. *Transport Policy*, 12(4):291–302.

Celli, F., Pianeai, F., Stilwell, D., and Kosinki, M. (2013). Workshop on computational personality recognition: Shared task. In *Proceedings of WCPR13, in conjunction with ICWSM-13*, Boston, July.

Cohen, E. (1972). Toward a sociology of international tourism. *Social Research*, 39(1):164–182.

Kasper, W. and Vela, M. (2011). Sentiment Analysis for Hotel Reviews. In *Proceedings of the Computational Linguistics-Applications Conference*. Polskie Towarzystwo Informatyczne, October.

Kessler, J. S., Eckert, M., Clark, L., and Nicolov, N. (2010). The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *4th Int'l AAAI Conference*

on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010).

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens et al., editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Mccrae, R. R. and Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52:81–90.

Nakayama, Y. and Fujii, A. (2013). Extracting Evaluative Conditions from Online Reviews: Toward Enhancing Opinion Mining. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 878–882, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pearce, P. (1988). *The Ulysses factor: Evaluating visitors in tourist settings*. New York, NY:Springer-Verlag.

Plog, S. (2001). Why destination areas rise and fall in popularity: an update of a cornell quarterly classic. *Cornell Hotel and Restaurant Administration Quarterly*, 42(3):13–24.

Salomon, I. and Mokhtarian, P. L. (1998). What happens when mobility-inclined market segments face accessibility-enhancing policies? Institute of transportation studies, working paper series, Institute of Transportation Studies, UC Davis.

Sayeed, A., Rusk, B., Petrov, M., Nguyen, H. C., Meyer, T. J., and Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities : LaTeCH ; proceedings of the workshop ; ACL HLT 2011 ; 24 June, 2011 Portland, Oregon, USA*, pages 69–77, Stroudsburg, PA. ACL.

Scheible, C. and Schütze, H. (2013). Sentiment relevance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 954–963, Sofia, Bulgaria, August. Association for Computational Linguistics.

Wiebe, J. and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. language resources and evaluation. In *Language Resources and Evaluation*, page 2005.

Wilson, T. and Wiebe, J. (2005). Annotating attributions and private states. In *Proceedings of ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

Ye, Q., Law, R., and Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182.

Ye, Q., Law, R., Gu, B., and Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2):634–639.

Yerva, S., Grosan, F., Tandrau, A., and Aberer, K. (2013). Tripeneer: User-based travel plan recommendation application. In *International AAAI Conference on Web and Social Media*.

Yoo, K. H. and Gretzel, U. (2011). Influence of personality on travel-related consumer-generated media creation. *Computers in Human Behavior*, 27(2):609–621.