Corpus-based Explorations of Affective Load Differences in Arabic-Hebrew-English

Carlo Strapparava[1] *Oliviero Stock*[1] *Ilai Alon*[2]

(1) FBK-irst, via Sommarive 18, Trento, Italy
(2) Department of Philosophy, Tel Aviv University, Israel
strappa@fbk.eu,stock@fbk.eu,ilaialon@post.tau.ac.il

ABSTRACT
This work is about connotative aspects of words, often not carried over in translation, which depend on specific cultures. A cross-language computational study is presented, based on exploitation of similarity techniques on large corpora of news documents in English, Arabic, and Hebrew. In particular, focus of the exploration is on specific terms expressing emotion, negotiation and conflict.

KEYWORDS: Multilinguality, Affective Language, Emotions in Language.

KEYWORDS IN L_2: Here a list of keywords in L_2 (if option used).

Proceedings of COLING 2012: Posters, pages 1201–1208,
COLING 2012, Mumbai, December 2012.

1 Introduction

Even an excellent human translator has problems in carrying over the target language all the culture-related aspects that go with words. If focus is put into emotion-related aspects the matter is even subtler. The relation of a word to emotion concepts may depend on ideology and in general on cultural aspects that can be inferred from extensive word usage rather than from what can be found in dictionaries. Of course it also depends on genres, different periods of text production, sociolinguistic characteristics of the text originators and so on.

In this paper, we describe a cross-language computational study based on exploitation of similarity techniques on large corpora of news documents in English, Arabic, and Hebrew. In particular, we focus our exploration on specific terms expressing emotion, negotiation and conflict.

Aside of the general scientific motivation, we had a specific motivation for starting this work: help overcoming unnecessary language problems in international negotiations involving different languages. In fact, perhaps the most damaging mistake in any negotiation is misunderstanding, especially that which is the result of ignorance and disregard. The need is to reduce one aspect of such misunderstanding.

During negotiations between Israelis and Palestinians for example, more than once the latter used the expression “the final solution" with reference to the Israeli-Palestinian conflict. For Israelis, as for many Westerners, this expression most importantly refers to the Holocaust. Thus, almost automatically it creates aversion, and is sometimes even interpreted as a threat. Or just consider the different valence of the word “honor" in an Arabic, English or Hebrew expression, particularly in an emotionally tense situation.

The aim of this work is to assess the emotional connotations of words which have more or less the same denotation in Arabic, Hebrew and English. Although Arabic and Hebrew have been studied for centuries by both Arab and foreign scholars, their emotive aspects have been rather neglected, at least from the semantic point of view. An exception among Arab scholars is Abdullah-T Shunnaq (Shunnaq, 1993). The view that emotions take part in the meaning of words was already made by McDougall during the Twenties’ of the last century (Gregg, 2005). (Ogden and Richards, 1923) and more recently (Kövecses, 2000) call attention on how emotions are treated in language. Davitz’s early work in the area of lexicography (Davitz, 1969) has recently gained greater interest with the advent of electronic media (Heise, 2001). (Kövecses, 2000) divides “emotion language" into expressive terms, terms literally denoting particular kinds of emotions, and figurative expressions, of which the latter “is the largest by far". On a similar line goes the cognitive approach of (Ortony et al., 1987).

On the other hand, cultures, and thus, languages, differ in the degree of emotionality, Arabic being considered high in this criterion (Shunnaq, 1993). This is even more evident for political terms, and in particular for those associated with conflict. In negotiation, recognition of the emotions of the other party is the first step on the road to conciliation. As (Irani, 1999) say “A first step in the process of healing, then, is the mutual acknowledgment by all parties of their emotions, viewpoints and needs.". On the negative side it has been said that “representing outcomes in affective terms leads to longer negotiation times and higher impasse rates" (Conlon and Hunt, 2002).

The important role of emotions in Middle East politics is also eloquently pointed to in an article by (Moïsi, 2007). In it he coined the phrase “clash of emotions", and argued that the Arab

world manifests a culture of humiliation. Others in the Middle East argue too, that the role of emotions is greater than that of civilizations in explaining violence in the region (Fattaha and Fierke, 2009). On the Social-personal level, emotion is closely tied to moral system of a culture, and thus plays a decisive role in communicating with that culture. As (Fattaha and Fierke, 2009) put it: “In this view, emotion finds expression only in a language and a culture, which is linked to a moral order and moral appraisal. In the Middle East, feelings are always “situated in configurations of interpersonal relationships." These are connected in turn with the honor-modesty system (honor, shame, and modesty) (Gregg, 2005).

Coming to us, as said, we had the goal of establishing a methodology and eventually reaching concrete results concerning the different connotations of corresponding terms in Arabic, Hebrew and English. For one of us the initial strategy was to proceed via questionnaires in Arabic, and Hebrew, with different populations.The initial attempt at getting results via questionnaires could not get very far, mainly because of small numbers. The subtlety of the questions and situations suggested crowdsourcing techniques were not appropriate as well. The idea then came of following a computational approach very much in line with the experience of the other two authors. In particular, we used corpus-based similarity techniques for exploring affective significance of words in different languages, with relevant practical implications.

2 Corpora and terms in focus

In the experiment of exploring similarity, we exploited three corpora in the respective languages.

Arabic: Arabic Gigaword Third Edition is a comprehensive archive of newswire text data acquired from Arabic news sources. The six distinct sources of Arabic newswire are: Agence France Presse, Assabah, Al Hayat, An Nahar, Ummah Press, and Xinhua News Agency. The total number of documents is about 1.500.000 in a span time from 1995 until 2007. The preprocessing on this corpus consisted of a conversion from Arabic to Buckwalter ascii encoding and of a postagging process with the AMIRA tool (Diab et al., 2004).

English: We collected about 400.000 Google-News in the years 2008/2009. The documents have been pos-tagged with the TextPro tool (Pianta et al., 2008).

Hebrew: We used a collection of news documents from three newspapers in the span time 1990 - 2002: Arutz7, The Marker, and HaAretz.The corpus includes 11.474 documents and it has been preprocessed with a pos-tagger (Itai and Wintner, 2008).

In building the datasets from the documents of the three corpora, we considered as parts of speech nouns, verbs, adjectives and adverbs.

In order to select a suitable set of terms of conflict and emotion terms, questionnaires were distributed among native speakers of Arabic and Hebrew respectively, i.e. students of universities (Tel Aviv, Haifa), colleges (al-Qasemi) and high-schools (Palestinian East Jerusalem). For English we felt it was not strictly necessary. Respondents were asked to provide words in the categories of emotion, conflict, conciliation and trust terms. Among the emotion terms, some would not be considered “emotions" by English speakers, but were still included by us. This method was employed in order to avoid contamination of the list by Western culture researchers (Wierzbicka, 1997), e.g. by only referring to the “universal" emotions, i.e., anger, fear, disgust, sadness,

Emotion Terms	
English	Frustration Respecting Contempt Faithfulness Humiliation Satisfaction Revulsion Security Taking-Interest Faith Abhorrence Tolerance Determination Extremism Empathy Mutual-Understanding Emergency Love Sadness Grudge Kindness Perplexion Fear Mercy Contentedness Fright Happiness Tenderness Friendship Weakness Persecution Compassion Violence Anger Fervor Amicability Hardheartedness Worry Subdue Power Hatred Pin Indifference Suffering Boredom Cordiality Despair Fondness Disgust
Arabic	تعاطف تطرف تصميم تسامح بغض ايمان اهتمام امان اشمئزاز ارتياح اذلال اخلاص احتقار احترام احباط ظلم ضعف صداقة شفقة سعادة سرور رعب رضا رحمة خوف حيرة حنان حقد حزن حب توتر تفاهم يأس مودة ملل معاناة محبة مبالاة ، عدم الم كره كراهية قوة قهر قلق قسوة فرح الفة غيرة غضب عنف عطف إشمئزاز ميل إلى
Hebrew	תעוב אמונה התעניינות בטחון סלידה נחת רוח השפלה נאמנות זלזול כבוד הסכול מבוכה רוך טינה עצב אהבה חרום הבנה הדדיות אמפתיה קיצוניות נחישות סובלנות אלימות חבה עוול חולשה חברות חמלה אושר שמחה חרדה שביעות רצון רחמים פחד סבל אדישות כאב שנאה שנאה כח הכנעה דאגה אכזריות ששון ידידותיות קנאות כעס נועל אהדה לבביות שעמום
Conflict Terms	
English	Racialism Coalition Innocent-people Respecting Fraternity Land Americanism Revenge-taking Degeneration Decline Humanism Solidarity Transfer Intimidation Clash-of-Civilizations Solidarity Normalization Cooperation Competition Expulsion Nationality Unlawful War Right-of-Return Blood Religion Peace Politics Struggle Zionism Oppressive Enmity Arab Secularism Globalization Racialism Killing Force Nationality Equality Muslim Confiscation Jews
Arabic	الترهيب ترانسفير ترابط الانسانية انحطاط انحطاط انتقام ارض اخوة احترام أبرياء ائتلاف عنصرية سياسة سلام دين دم حق العودة حرب حرام جنسية تهجير تنافس تعاون التطبيع تضامن تصادم الحضارات مساواة قومية قوة قتل عنصرية عولمة علمانية عربي العداوة - آخر شيء في النزاع ظالم صهيزنية صراع يهود مصادرة مسلم
Hebrew	שקיעה התנקמות אמריקאיות אדמה אחווה כבוד חפים מפשע קואליציה נועזות נורמליזציה סולידריות ההתנגשות היות טרור טראנספר סולידריות אנושיות מדיניות שלום דת דם זכות השיבה מלחמה לאומיות נרוש תחרות שתוף פעולה לאומיות כח הרג נועזות גלובליזציה חילוניות ערבי איבה עושק ציונות מאבק יהודים החרמה מסלם שויון
Conciliation Terms	
English	Compromise Concessions Conciliation Negotiating Deal
Arabic	صفقة أخذ وعطاء تصالح تنازلات تسوية
Hebrew	עסקה משא ומתן פיוס ויתורים פשרה
Trust Terms	
English	Double-cross Betrayal Treason Loyalty Confidence Trust Deceit Credibility Treachery Reliability Fraud
Arabic	غش إخلاص خيانة مصداقية خدعة ثقة أمانة وفاء خيانة غدر خداع
Hebrew	מעל מהימנות בגידה נאמנות אמינות רמאות אמון נאמנות בגידה בגידה הונאה

Table 1: Emotion, conflict, conciliation, and trust terms in the three languages

happiness, surprise. The terms that emerged as important in the questionnaires in Arabic and Hebrew were in the focus list of the computational experiment. Their translations (selected by a human expert) in the two other languages were picked out as well. In Table 1 the terms used in our experiments are reported.

Frustration	*Land*	0.311	*Anger*	*Politics*	0.376	*Extremism*	*Zionism*	0.101
احباط	ارض	0.640	غضب	سياسة	0.341	تطرف	صهيزنية	0.316
תסכול	אדמה	0.184	כעס	מדיניות	0.132	קיצוניות	ציונות	0.157
Mercy	*Respecting*	0.149	*Fear*	*Double-cross*	0.154	*Extremism*	*Arab*	0.114
رحمة	احترام	0.021	خوف	خداع	0.691	تطرف	عربي	0.068
רחמים	כבוד	0.500	אהבה	הונאה	0.059	קיצוניות	ערבי	0.404
Hatred	*Fraud*	0.057	*Fright*	*Double-cross*	0.305	*Extremism*	*Blood*	0.029
كراهية	غش	0.209	رعب	خداع	0.645	تطرف	دم	0.261
שנאה	מעל	0.325	חרדה	הונאה	0.001	קיצוניות	דם	0.092
Sadness	*War*	0.074	*Anger*	*Double-cross*	0.105	*Extremism*	*Intimidation*	0.297
حزن	حرب	0.096	غضب	خداع	0.717	تطرف	الترهيب	0.436
עצב	מלחמה	0.209	כעס	הונאה	0.150	קיצוניות	טרור	0.085
Fright	*Killing*	0.078	*Fright*	*Globalization*	0.089	*Love*	*Zionism*	0.045
رعب	قتل	0.545	رعب	عولمة	0.224	حب	صهيزنية	0.057
חרדה	הרג	0.220	חרדה	גלובליזציה	0.016	אהבה	ציונות	0.237
Fear	*Politics*	0.366	*Fright*	*Confiscation*	0.008	*Love*	*Arab*	0.025
خوف	سياسة	0.330	رعب	مصادرة	0.250	حب	عربي	0.247
פחד	מדיניות	0.079	חרדה	החרמה	0.064	אהבה	ערבי	0.056

Table 2: Some similarity values in the three corpora

3 Technique

As a corpus-based measure of semantic similarity we exploited latent semantic analysis (LSA) proposed by Landauer (Landauer et al., 1998). In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix **T** representing the corpus.

SVD is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find correlations among its rows and columns. In our case, SVD decomposes the term-by-document matrix **T** into three matrices $\mathbf{T} = \mathbf{U}\Sigma_k\mathbf{V}^T$ where Σ_k is the diagonal $k \times k$ matrix containing the k singular values of **T**, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k$, and **U** and **V** are column-orthogonal matrices. When the three matrices are multiplied together the original term-by-document matrix is re-composed. Typically we can choose $k' \ll k$ obtaining the approximation $\mathbf{T} \simeq \mathbf{U}\Sigma_{k'}\mathbf{V}^T$.

LSA can be viewed as a way to overcome some of the drawbacks of the standard vector space model (sparseness and high dimensionality). In fact, the LSA similarity is computed in a lower dimensional space, in which second-order relations among terms and texts are exploited. The similarity in the resulting vector space is then measured with the standard *cosine* similarity. Note also that LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, word sets, and texts. It is possible to represent set of words in the semantic space using the *pseudo-document* text representation for LSA computation, as described by Berry (Berry, 1992). In practice, each text segment is represented in the LSA space by summing up the normalized LSA vectors of all the constituent words, using also a *tf.idf* weighting scheme. For the experiments reported in this paper, we run the SVD operation respectively on the three preprocessed corpora described in the previous section, using $k' = 400$ dimensions.

4 Results and discussion

To give an idea about different behaviors of corresponding terms in the three languages, in Table 2 we report some similarity values. In the initial part of the list we show similarity measures between emotion terms and some generic terms. The entries that follow in the list include more opinionated terms. The differences among values in the three languages are quite noticeable and can be considered as evidence of different sociocultural perceptions of the involved terms.

These results suggest that the proposed techniques are a viable tool for approaching cultural differences that emerge in different languages.

Of course, in the future, more specialized and, when possible, strictly aligned corpora can be used for the involved languages, as the applied context may require.

The computational approach we have presented has proven to be very promising: looking at specifically critical words for a sensitive situation like a multilingual negotiation in a bitter conflict, different emotional connotations of words, which are considered as the right translation, tend to appear clearly. From the applied point of view we are taking into consideration the development of an interface that would offer a quick perception of these different connotations across the involved languages, yielding an immediate feeling of the emotional aspect often lost in translation.

Acknowledgments

We thank Shuly Wintner, Noam Ordan, and Yulia Tsvetkov for providing and preprocessing the Hebrew corpus, and Arianna Bisazza for her hints regarding Arabic preprocessing. Carlo Strapparava was partially supported by Eurosentiment FP7 EU-project.

References

Berry, M. (1992). Large-scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6(1).

Conlon, D. and Hunt, C. (2002). Dealing with feeling: the influence of outcome representations on negotiation. *International Journal of Conflict Management*, 13(1):38–58.

Davitz, J. R. (1969). *The Language of Emotion*. Academic Press.

Diab, M., Hacioglu, K., and Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Short Papers*, pages 149–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Fattaha, K. and Fierke, K. (2009). A clash of emotions: The politics of humiliation and political violence in the middle east. *European Journal of International Relations*, 15(1):67–93.

Gregg, G. S. (2005). *The Middle East: A Cultural Psychology*. Oxford University Press.

Heise, D. R. (2001). Project magellan: Collecting cross-cultural affective meanings via the internet. *Electronic Journal of Sociology*.

Irani, G. E. (1999). Islamic mediation techniques for middle east conflicts. *Middle East Review of International Affairs*, 3(2).

Itai, A. and Wintner, S. (2008). Language resources for hebrew. *Language Resources and Evaluation*, 42(1):75–98.

Kövecses, Z. (2000). *Metaphor and Emotion: Language, Culture, and Body in Human Feeling*. Cambridge University Press.

Landauer, T. K., Foltz, P., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25.

Moïsi, D. (2007). The clash of emotions-fear, humiliation, hope, and the new world order. *Foreign Affairs*, 86.

Ogden, C. K. and Richards, I. A. (1923). *The Meaning of Meaning*. Harcourt, Brace & World.

Ortony, A., Clore, G. L., and Foss, M. A. (1987). The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, 53:751–766.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The TextPro tool suite. In *Proceedings of 6th edition of the Language Resources and Evaluation Conference (LREC)*.

Shunnaq, A. (1993). Lexical incongruence in arabic-english translation due to emotiveness in arabic. *Turjuman*, 2:237–263.

Wierzbicka, A. (1997). *Understanding cultures through their key words.* Oxford University Press.