Interaction Mining: the new frontier of Call Center Analytics

Vincenzo Pallotta[1], Rodolfo Delmonte[1,2], Lammert Vrieling[1], David Walker[1]

Interanalytics
Rue des Savoises, 19 1205 Geneva, Switzerland
<firstname.surname>@interanalytics.ch

Department of Computational Linguistics
University 'Ca Foscari Venice, Italy.
delmont@unive.it

Abstract In this paper, we present our solution for pragmatic analysis of call center conversations in order to provide useful insights for enhancing Call Center Analytics to a level that will enable new metrics and *key performance indicators* (KPIs) beyond the standard approach. These metrics rely on understanding the dynamics of conversations by highlighting the way participants discuss about topics. By doing that we can detect situations that are simply impossible to detect with standard approaches such as controversial topics, customer-oriented behaviors and also predict customer ratings.

1 Introduction

Call centers data represent a valuable asset for companies, but it is often underexploited for business purposes. By call center data we mean all information that can be gathered from recording calls between representatives (or agents) and customers during their interactions in call centers. These interactions can happen over multiple different channels including telephone, instant messaging, email, web forms, etc. Some information can be collected without looking at the content of the interaction, by simply logging the system used for carrying the conversation. For example, in call centers, *calls duration* or *number of handled calls* can be measured by software for telephony communication. We call these measures standard call center *Key Performance Indicators* (KPIs). With standard KPIs, only limited analytics can be done providing a partial understanding of the call center performance and no information whatsoever is collected about what is going on within the interaction. *Call Center Analytics* is aimed at solving the above issue by enabling tapping into the content of conversations. The technology for Call Center Analytics is still in its infancy and related commercial products have not yet achieved maturity. This is due to two main factors: i) it is highly dependent on quality of speech recognition technology and ii) it is mostly based on text-based content analysis.

Our approach to Call Center Analytics is based on *Interaction Mining*, a new research field aimed at extracting useful information from conversations. In contrast to Text Mining (Feldman and Sanger 2006), Interaction Mining is more robust, tailored for the conversational domain, and slanted towards *pragmatic* and *discourse* analysis. In particular, with our approach we were able to achieve the following four objectives:

1. Identify Customer Oriented Behaviors, which are highly correlated to positive customer ratings (Rafaeli et al. 2007);
2. Identify Root Cause of Problems by looking at controversial topics and how agents are able to deal with them;
3. Identify customers who need particular attention based on history of problematic interactions;
4. Learn best practices in dealing with customers by identifying agents able to carry cooperative conversations. This knowledge coupled with customer profiles can be used effectively in intelligent skill-based routing[1]

The article is organized as follows: in section 2 we review current Speech Analytics technology and make the case for Interaction Mining approach in order to address the current business challenges in call centers quality monitoring and assessment. In section 3 we present our Interaction Mining solution based on a specific kind of pragmatic analysis: the Argumentative Analysis and its implementation with the A3 algorithm. In section 4 we showcase our solution for call center analytics and the implementation of new relevant metrics and KPIs for call center quality monitoring. We conclude the article with a discussion on the achieved results and a roadmap for future work.

2 Call Center Analytics Needs Interaction Mining

Call center data contain a wealth of information that usually remains hidden. Key Performance Indicators (KPIs) for call centers performance can be classified into three broad categories (Baird 2004):

1. Agent Performance Statistics: these include metrics such as *Average Speed of Answer*, *Average Hold Time*, *Call Abandonment Rate*, *Attained Service Level*, and *Average Talk Time*. They are based on quantitative measurements that can be obtained directly through ACD[2] Switch Output and Network Usage Data.

[1] http://en.wikipedia.org/wiki/Skills-based_routing

[2] Automatic Call Distribution.

2. Peripheral Performance Data: these include metrics such as *Cost Per Call*, *First-Call Resolution Rate*, *Customer Satisfaction*, *Account Retention*, *Staff Turnover*, *Actual vs. Budgeted Costs*, and *Employee Loyalty*. These metrics are mostly quantitative, with the exception of *Customer Satisfaction* that is usually obtained through Customer Surveys.
3. Performance Observation: these include metrics such as *Call Quality*, *Accuracy and Efficiency*, *Adherence to Script*, *Communication Etiquette*, and *Corporate Image Exemplification*. These are qualitative metrics based on analysis of recorded calls and session monitoring by a supervisor.

Minnucci (2004) reports that the most required metrics by call center managers are indeed the qualitative ones topped by Call Quality (100%) and Customer Satisfaction (78%). However, these metrics are difficult to implement with the adequate level of accuracy[3]. Most call center quality monitoring dashboards[4] implementing standard metrics are now only able to display information related to service-level measures (Agents and Peripheral Performance data), namely how fast and how many calls agents able to handle. Because of recent improvements of speech recognition technology (Neustein 2010), Speech Analytics is viewed as a key element for implementing call center quality monitoring. As pointed out by Gavalda and Schlueter (2010), Speech Analytics is becoming "*an indispensable tool to understand what is the driving call volume and what factors are affecting agents' rate of performance in the contact center*."

2.1 Interaction Mining

Interaction Mining is an emerging field in Business Analytics that contrasts the standard approach based on Text Mining (Feldman and Sanger 2006). In Text Mining the assumption made is that input is textual and can be treated as sets of content-bearing terms. This assumption is no longer valid in conversational input. Non-content words such as conjunctions, prepositions, personal pronouns and interjections are extremely important in conversations cannot be filtered out as they bear most of their pragmatic meaning. As pointed out in Pallotta et al. (2011) there are several advantages of moving to Interaction Mining for generating intelligence from conversational content. It is important to note that while the purpose is similar, namely turning unstructured data into structured data for performing quantitative analysis, Text Mining focuses on pattern extraction from *documents*. This is no longer the case with conversational content as the units of information in conversational content are *dialogue turns* and typically they are significantly shorter than documents. This means that the input has to be fully linguistically processed in order to understand its *pragmatic function* in the conversation. For instance, a

[3] Accuracy is defined in (Baird 2004) as true indication and it depends on the actual level of performance attainment, especially with regard to statistical validity.

[4] An example of call center analytics dashboard is available at: http://demos7.dundas.com/HVR.aspx

simple turn containing just one single word like “Yes” or “No” can make a substantial difference in the interpretation of a whole conversation.

Interaction Mining tools are substantially different than those employed in Text Mining. Machine learning approaches are no longer a viable option since data are very sparse and attempts have failed in providing satisfactory results so far (Rienks and Verbree 2006; Hakkani-Tür 2009). In Pallotta et al (2011) we have provided evidences that bag-of-words approach simply is not suitable for pragmatic indexing of conversations, and therefore useless for tasks as Question Answering or Summarization. Another limitation of Text Mining is in Sentiment Analysis. As we have previously shown in Delmonte and Pallotta (2011), shallow linguistic processing and machine learning often provide misleading results. Therefore, we advocated for a deep linguistic understanding of input data even for standalone contributions such as product reviews. In Interaction Mining, Sentiment Analysis issues become even more compelling because sentiment about a topic is not fully condensed in a single turn but it develops along the whole conversation. For example, it is very common that dissent is expressed toward other the opinion of other participants in the conversation rather than to the topic under discussion. Sentences like “why do you think product X is bad?” would be simply mistakenly considered as a negative attitude to product X in a bag-of-word approach.

2.2 Related Work

Current approaches to Call Center Analytics are mostly based on Speech Analytics, which is fundamentally based on Search and Text Mining technology. Recorded speech is phonetically indexed and searched: *phonetic transcription* is more reliable and accurate than Large-Vocabulary Continuous Speech Recognition (LVCSR) and keyword queries can easily be turned into its phonetic counterpart for search. With this approach one can search for occurrence of specific words in calls. Its simplicity is at the same time its strength and weakness. On the one hand the method is fast and accurate but, on the other hand, it is limited to its applicability for generating adequate insights on calls because the *context* of word’s occurrence is lost and it can only recovered by physically listening to the audio excerpt where the searched word occurs.

While still very high compared to human performance, the Word Error Rate (WER) [5] of LVCSR systems shows a promising trend as reported by the NIST Speech-To-Text Benchmark Test History 1988-2007 (Fiscus et al. 2008). Instead of downgrading the analysis capabilities we believe it is more appropriate to make the analysis less sensitive to WER. In other words, we want a robust solution capable of delivering approximate but still sound measurements for content-based metrics. We will show in the next sections that our approach to Interaction Mining is robust and it can properly deal with output from LVCSR systems.

---

[5] http://en.wikipedia.org/wiki/Word_error_rate.

Another common approach to the analysis of call center data is that of automatic call categorization through supervised machine learning (Gilman et al. 2004; Zweig et al. 2006; Takeuchi et al. 2009). These methods failed in providing satisfactory results even in very broad categories. The problem lies on the fact that data is very sparse and that huge amount of training data is necessary to achieve reasonable discriminatory power.

Unsupervised learning provides better results for domain-specific classes as shown in Tang et al. (2003). However, the sensibility to domain represents a big issue. Moreover, this type of categorization – i.e. topics of calls – helps little to understand if a call is satisfactory or not. It might be better suited for retrieval and aggregation of other quality-oriented information.

3 Argumentative Analysis for Interaction Mining

Our approach to pragmatic analysis for Interaction Mining is rooted on *argumentative analysis* (Pallotta 2006). Argumentation is a pervasive pragmatic phenomenon in conversations. Purposeful conversations are very often aimed at reaching a consensus for a decision or to negotiate opinions about relevant topics.

The argumentative structure defines the different patterns of argumentation used by participants in the dialog, as well as their organization and synchronization in the discussion. From this perspective, we adopted in (Pallotta 2006; Pallotta et al. 2007) an argumentative coding scheme, the Meeting Description Schema (MDS). In MDS, the argumentative structure of a meeting is composed of a set of topic discussion episodes, where several issues are discussed through the proposal of alternatives, solutions, opinions, ideas, etc. in order to achieve a satisfactory decision. Proposals can be accepted or challenged through acts of rejecting or asking questions.

3.1 Automatic Argumentative Annotation

The core of our solution is a system that extracts the argumentative structure of conversations. This system is based on adapting and extending the GETARUNS text understanding system (Delmonte 2007; 2009). Details of the *Automatic Argumentative Annotation* (A3) algorithm are available in Delmonte et al. (2010). The system has been evaluated on manually transcribed conversations from the ICSI meeting corpus (Janin et al. 2001) and annotated by Pallotta et al. (2007). With a Recall of 97.53%, we computed the Precision as the ratio between the number of Correct Argumentative Labels and the number of Argumentative Labels Found, which corresponds to 81.26%. The F-score is 88.65%.

In order to check the robustness of the A3 algorithm when applied to automatically transcribed conversations, we evaluated the A3 algorithm on similar data that were transcribed using a state-of-the-art LVCSR system (Fiscus et al. 2008). We have measured the performance of our system and observed a degradation of only 11.7% of the overall performance with a LVCSR system showing an average WER of 30% (Hain et al. 2009). These results are quite promising and, coupled with expected improvements in LVCSR technology and further tuning of the system, they provide us with a solid basis for development.

3.2 Multi-word expressions

One key issue with conversation is that topics are not expressed by single words but very often by compounds. Hence, quality of topic detection can be improved if the lexicon contains domain-specific multi-word expressions. We thus run a multi-word expressions extraction tool (Seretan and Wehrli. 2009) to identify the most frequent compounds in the corpus and compare them with the topics detected by the GETARUNS system. The top 10 extracted multi-word expressions[6] and topics are shown in Table 1.

Multi Word Expression	Score
1. Calling Chase	475,4809
2. Account number	300,2746
3. Gross balance	282,5876
4. Direct deposit	247,4588
5. Savings account	189,3173
6. And available	186,6647
7. Social security number	159,8058
8. Area code	146,8807
9. Daytime phone number	143,3286
10. Most recent	126,4333

a

Topic	% of total
1. Chase	5,26%
2. Social security number	3,41%
3. Checking account	2,75%
4. Moment	2,33%
5. Statement	2,16%
6. Money	2,00%
7. Savings account	1,45%
8. Dollar	1,37%
9. Days	1,35%
10. Phone number	1,23%

b

Table 1. Multi Word Expression extracted from the corpus (a) and GETARUN topics (b).

While there is a predictable overlapping it is interesting to see that some domain-dependent terms were detected by the multi-word extraction system but they were not included in the lexicon of our system such as “gross balance” and “and available”, and “direct deposit”[7]. Enriching the lexicon with these terms would greatly improve the pragmatic analysis of conversations.

[6] The score for multi-word expression represents the log-likelihood ratio statistics representing the association strength between the component words (Dunning 1993).

[7] The “and” and “Available” are detected as a multi-word expression because they occur frequently in the corpus as the pattern “Gross and Available balance”.

4 Experiments with Call Center Data

In this section we present the results of an experiment where we applied our Interaction Mining technology to actual call center data. The goal was to find out if the argumentative analysis, coupled with sentiment analysis, could indeed provide useful insights for Call Center Analytics. The results show, in particular, that we were able to achieve the objectives we introduced in Section 1 and therefore implement the relevant and most requested KPIs in call center quality management.

In our experiment we used a corpus of 213 manually transcribed conversations of a help desk call center in the banking domain. Each conversation has an average of 66 turns and an average of 1.6 calls per agent. This corpus was collected for a study aimed at identifying conversational behaviors that could favor satisfactory interaction with customers (Rafaeli et al. 2007).

Customer Oriented Behaviors	
anticipating customers requests	22,45%
educating the customer	16,91%
offering emotional support	21,57%
offering explanations / justifications	28,57%
personalization of information	10,50%

Table 2. Customer-Oriented Behaviors from Call Center data (Rafaeli et al. 2007)

Table 2 shows the identified COBs and their distribution in the data. Unfortunately, only a very small portion of the data (2.5%) was manually annotated with COBs, which prevented us from either performing a statistically sound correlation study or train a model.

4.1 Argumentative analysis of call center conversations

We run our A3 algorithm to the call center data and we visualized the results with off-the-shelf business intelligence tools. We used Tableau 6.0[8], which revealed to be a suitable tool for getting insightful multi-dimensional aggregations and charts into dashboards for addressing the four quality monitoring objectives mentioned in the beginning of this section.

Identify Customer Oriented Behaviors

We noticed that COBs showed a high resemblance to our argumentative categories and that they might correlate as well.

[8] http://www.tableausoftware.com

As shown in Fig. 1 the "Provide Explanation/Justification" and "Suggest" categories highly correlate with COBs. Combined with additional extracted information such as Sentiment and Subjectivity (see Pallotta and Delmonte (2011) for more details), we can safely conclude that COBs can be easily predicted by our system.

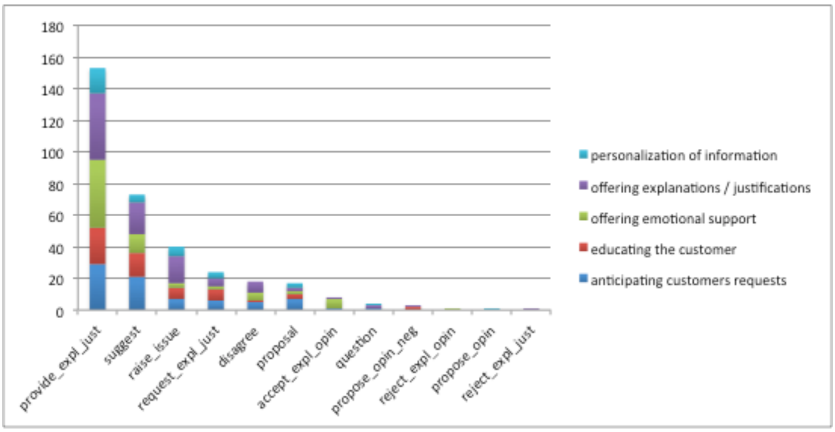

Fig. 1. Correlation between argumentation and customer-oriented behaviors**Identify Root Cause of Problems**

By looking at *controversial topics* we can identify root cause of problems in call centers. We selected the worst 20 topics ranked according to frequency of negative attitudes obtained by the Sentiment Analysis module. Fig. 2. shows a dashboard that can be used to detect controversial topics and thus help in spotting unsolved issues.

Fig. 2. Problem spotting dashboard

The cooperativeness score is a measure obtained by averaging the score obtained by mapping argumentative labels of each turn in the conversation into a [-5 +5] scale. The mapping is shown in Table 3.

Argumentative Categories	Level of Cooperativeness
Accept explanation	5
Suggest	4
Propose	3
Provide opinion	2
Provide explanation or justification	1
Request explanation or justification	0
Question	-1
Raise issue	-2
Provide negative opinion	-3
Disagree	-4
Reject explanation or justification	-5

Table 3. Mapping table for argumentative categories to levels of cooperativeness

The mapping is hand crafted and rooted on Bales's Interaction Process Analysis framework (Bales, 1950), where uncooperativeness (i.e. negative scores) is linked to high level of criticism, which is not balanced by constructive contributions (e.g. suggestions and explanations). This mapping provides a reasonable indicator of controversial (i.e. uncooperative) conversations.

The dashboard in Fig. 3 highlights the top 10 most discussed topics and the level of cooperation of the discussions. In the main pane, rows correspond to speakers and for each topic a square is displayed whose dimension represents the number of turns and the color its cooperativeness score. The histograms show the overall cooperativeness scores.

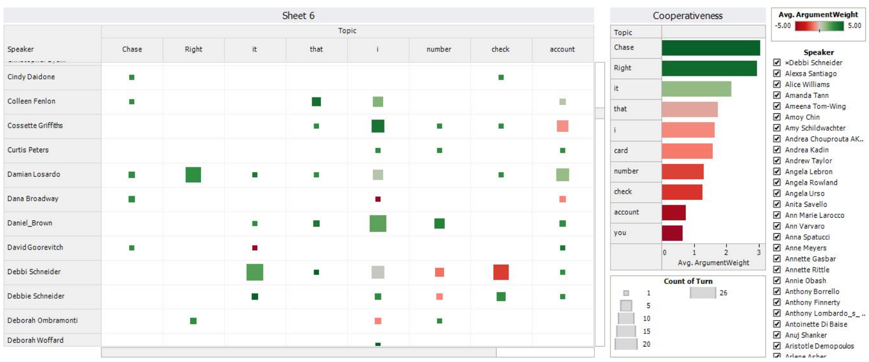

Fig. 3 Topic and Behaviors Dashboard

Identify problematic customers

A critical issue in this domain is that customers are not all the same and need to be treated differently according to their style of interaction. There are agents with interpersonal skills who are able to comfortably deal with demanding customers. Agents who show consistently positive cooperativeness can be assumed to be more suitable to deal with extreme cases. Customers who have already shown negative or uncooperative attitudes could be routed to more skilled agents in order to maximize the overall call center performance (i.e. customer satisfaction).
We present a dashboard in Fig. 4. where problematic customers can be identified and given a particular care.

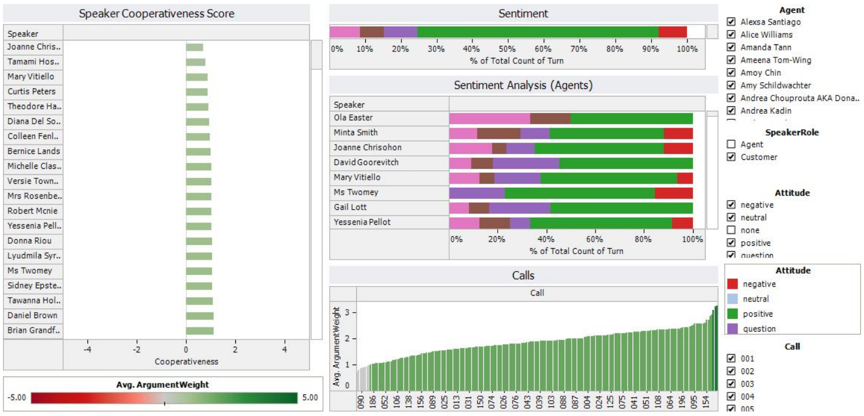

Fig. 4. Speakers Assessment Dashboard

With this dashboard speakers (agents or customers) are ranked according to their cooperativeness score. In the right-hand pane, also the sentiment analysis results are displayed and compared to the overall sentiment score. The analyst can then drill through a specific customer and visualize a specific customer and the calls he/she made.

Learn best practices from conversations

This objective results from all the insights gained through the presented visualizations. In particular, Fig. 4. with Agent filtering activated allows one to visualize overall and specific agent's behavior. Best scoring agents can be taken as models and their interaction used as models. While most of available solutions for skill-based inbound call routing are based on ACD information such as area codes for agent's language selection or based on IVR[9] for option selection. Additionally, the

[9] Interactive Voice Response

agent selection is often based on efficiency measures in order to optimize the costs and workload (e.g. by assigning the fastest agent to the longest queue). If this strategy might maximize efficiency, they are insufficient to maximize customer satisfaction. We advocate for skill-based call routing based on interpersonal qualities and by influencing the agent selection by cooperativeness requirements.

5 Conclusions

In this article we have presented a new approach to Call Center Analytics based on Interaction Mining, contrasting Text Mining, which is currently used in Speech Analytics. We presented an Interaction Mining tool for pragmatic analysis of conversations based on argumentation theory. We showed that our system is robust enough to deal with automatically transcribed speech, as it would be the case in Call Center Analytics. We conducted an evaluation the impact of this technology to a real case by applying our tools to a dataset of call center conversations in the banking domains. We presented the extracted information in several dashboards with the goal of implementing relevant KPIs for Call Center Analytics.

As for future work we would like to explore other pragmatic dimensions beyond argumentation. This might be relevant in the Call Center domain to look at COBs that are more related to emotional support or providing personalized information, which do not relate directly with argumentation. We need to consider finer granularity in argumentative analysis, for instance at clause level. This might be helpful when a single turn carries several argumentative functions. This would definitively improve the quality of the analysis. Our goal is to implement other KPIs for the Call Center domain such as adherence to scripts and corporate image exemplification. In order to achieve these challenging objectives, new types of pragmatic analysis will be required. Finally, we would like to explore the possibility of automatically learning agent and customer profiles from our analysis in order to implement more effective skill-based call routing.

6 References

Baird H. (2004). Ensuring Data Validity Maintaining Service Quality in the Contact Center. Telecom Directions LLC.

Delmonte R. (2007). *Computational Linguistic Text Processing – Logical Form, Semantic Interpretation, Discourse Relations and Question Answering*, Nova Science Publishers, New York.

Delmonte R. (2009). *Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution*, Nova Science Publishers, New York.

Delmonte R., Bristot A. and Pallotta V. (2010). Deep Linguistic Processing with GETARUNS for spoken dialogue understanding. In *Proceedings of LREC 2010*. Malta 18-23 May, 2010.

Delmonte R. and Pallotta V. (2011). Opinion Mining and Sentiment Analysis Need Text Understanding. In Pallotta V., Soro A., Vargiu E. (eds.). *Advances in Distributed Agent-Based Retrieval Tools*. Series: Studies in Computational Intelligence, Springer Verlag.

Dunning T. (1993). Accurate methods for the statistics of surprise and coincidence , Computational Linguistics, 19(1): 61-74.

Feldman, R., and Sanger, J. (2006). The Text Mining handbook. Advanced approaches in analyzing unstructured data. Cambridge University Press.

Fiscus J.G., Ajot J. and Garofolo J.S. (2008). The Rich Transcription 2007 Meeting Recognition Evaluation. In Stiefelhagen R., Bowers R. and Fiscus J. (Eds.) Multimodal Technologies for Perception of Humans. Lecture Notes In Computer Science, Vol. 4625. Springer-Verlag, Berlin, Heidelberg. pp. 373-389.

Hakkani-Tür D. (2009). Towards automatic argument diagramming of multiparty meetings. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Taipei, Taiwan. April, 2009.

Gilman A., Narayanan B. and Paul S. (2004). Mining call center dialog data. In Zanasi A., Ebecken N.F.F. and Brebbia C.A. (Eds.) *Data Mining V*. WIT Press.

Hain T., Burget L., Dines J., Garner P.N., El Hannani A., Huijbregts M., Karafiat M., Lincoln M. and Wan V. (2009). The AMIDA 2009 Meeting Transcription System In *Proceedings of INTERSPEECH 2009*, Makuhari, Japan, pp. 358-361.

Janin A., Baron D., Edwards J., Ellis D., Gelbart D., Morgan N., Peskin B., Pfau T., Shriberg E., Stolcke A. and Wooters C. (2003). The ICSI Meeting Corpus. In *Proceedings of IEEE/ICASSP 2003*, 6-10 April 2003, Hong Kong, vol. 1, pp. 364-367.

Minnucci J. (2004). Call Center KPIs: A Look at How Companies Are Measuring Performance, a special report published by ICMI Inc., 2004, p. 2.

Neustein A. (ed.) (2010). *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, DOI 10.1007/978-1-4419-5951-5_10, Springer , LLC.

Pallotta V. (2006), Framing Arguments. In *Proceedings of the International Conference on Argumentation* ISSA, June 2006, Amsterdam, Netherlands.

Pallotta V., Seretan V. and Ailomaa M. (2007). User requirements analysis for Meeting Information Retrieval based on query elicitation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (ACL 2007), pages 1008–1015, Prague.

Pallotta V., Vrieling L. and Delmonte R. (2011). Interaction Mining: Making business sense of customers conversations through semantic and pragmatic analysis. In Zorrilla M., Mazón J-N., Ferrández Ó., Garrigós I., Daniel F. and Trujillo J. (Eds.) *Business Intelligence Applications and the Web: Models, Systems and Technologies*. IGI Global to appear.

Rafaeli, A., Ziklik, L., and Doucet, L. (2007). The impact of call center employees' customer orientation behaviors on customer satisfaction. *Journal of Service Research*.

Rienks R. and Verbree D. (2006). About the usefulness and learnability of argument-diagrams from real discussions. In *Proceedings of the 3rd International Machine Learning for Multimodal Interaction Workshop* (MLMI 2006), May 1-4, 2006, Bethesda (MD), USA.

Seretan, V. and Wehrli, E. (2009). Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, 43(1), 71–85.

Tang M., Pellom, B. and Hacioglu, K. (2003). Call-type classification and unsupervised training for the call center domain. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding ASRU '03*, pp. 204 – 208.

Takeuchi H., Subramaniam L.V., Nasukawa T. and Roy S. (2009). Getting insights from the voices of customers: Conversation mining at a contact center. *Information Sciences* n. 179 pp. 1584–1591, Elsevier.

Zweig G., Siohan O., Saon G., Ramabhadran B., Povey D., Mangu L. and Kingsbury B. (2006). Automated Quality monitoring for call centers using speech and NLP technologies. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume*, pages 292–295, New York City, Association for Computational Linguistics.