Analysis and Models of Language in Electronic Negotiations

Marina Sokolova Vivi Nastase Stan Szpakowicz
Mohak Shah
SITE, University of Ottawa, Ottawa, Canada
{sokolova,vnastase,szpak,mshah}@site.uottawa.ca

Abstract

In electronic negotiations, language is the negotiators' principal means of reaching a deal. They use language to persuade, threaten and query, aiming to get the largest piece of the pie. An electronic negotiation support system has been gathering textual data. We study these data to build a model of the language used in e-negotiations and to observe, through language patterns, the behaviour of negotiators.

1 Motivation

Negotiation is a special, and quite interesting, type of interpersonal communication. It occurs in business and casual settings. Electronic negotiations (e-negotiations), conducted by email or other electronic means [11], are a relatively new phenomenon. They can be categorized both as business communication and as computer-mediated communication (CMC) [2, 6].

In the absence of nonverbal communication in e-negotiations (such as body language), text messages and non-textual data (offers) that negotiators exchange are the only source of behaviour disclosure [5]. In most e-negotiation systems negotiators exchange free-form messages. Non-textual data is exchanged through negotiation support systems(NSSs) and comes in different forms, depending on the system.

We develop a general analysis framework that can apply to data collected by any NSS. Among the types of data collected by NSSs, we concentrate on textual data, that is, messages exchanged; most systems

collect such messages. We study the language used in text messages exchanged via a Web-based NSS by negotiators from many countries.

We first compare the language used in e-negotiations with language used in other genres (news articles, literature, casual conversations). We show that the collection of messages exchanged during negotiations differs from any other corpus analyzed using Natural Language Processing (NLP) methods. The negotiators come from different countries, they have different professional and educational background, and their command of English varies. This makes e-negotiation data an interesting artifact.

The purpose of studying language in the context of negotiations is manifold. One goal is to find commonalities in linguistic behaviour of people brought together by a common activity but coming from different backgrounds and cultures. To address this task, we develop a methodology based on statistical and analytical methods. We observe that different subsets of negotiators share interesting characteristics. These subsets represent either negotiators who play a specific role (buyer or seller), or negotiators who have participated in a specific type of negotiation (completed or uncompleted).

Another goal is to use language to analyze behaviour. Because there is no communication using nonverbal means (for example body language), messages and non-textual data, when they exist, carry all intentions of their author. If the intention is to threaten, promise, or argue, then this should be evident from the message and additional data. We aim to find linguistic expressions that convey different types of strategies [5] that negotiators use, and to find which combinations result in a successful completion of the negotiation process.

This paper focuses on our first goal. We compare our data with various corpora used in NLP research. We use statistical methods to find what makes our data unique, and what common characteristics are shared by the negotiators involved in the recorded negotiations. This paper presents two different data models. To find quantitative characteristics of data we build a statistical language model. To find qualitative characteristics of data we build a lexicon that contains syntactic and semantic information for each word used. Both models are used to classify e-negotiation data [16].

This study continues the research on the language of negotiations [17], part of an on-going major project [9].

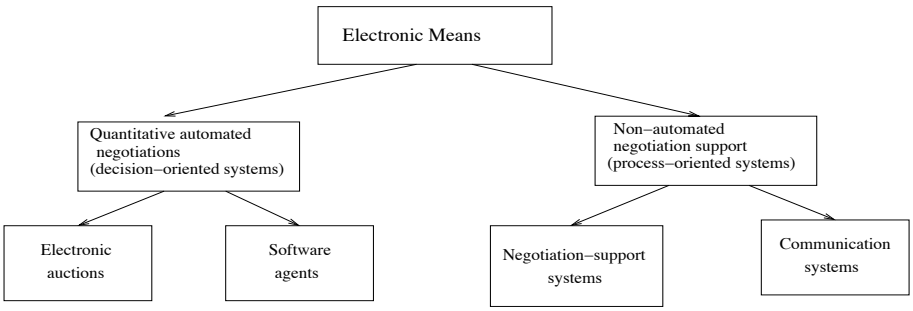

Figure 1: Types of Electronic Means in E-negotiations

Table 1: Types of Data Gathered by Process-Oriented Systems

Systems	Communication	Negotiation-support	
		Communication	Decision-support
Data	messages, offers	messages, offers	pre-negotiation phase data, negotiation phase data, post-negotiation phase data

2 E-negotiation Systems

Automated e-negotiation systems, such as electronic auctions or intelligent negotiation agents, make autonomous decisions. Other support systems leave decision-making to the users, but provide a variety of analytical tools [15]. This is shown in Fig. 1. We concentrate on the type of data gathered by NSSs, which combine decision support with electronic communications [9].

Data coming from electronic negotiations depend on the features of the electronic means that support the negotiations [9]. First consider systems that facilitate communication. If this allows only the exchange of free-form messages [18], the data will resemble a collection of email dialogues. In some cases the electronic means provide templates for writing messages [15], but the small volume of such data excludes statistical NLP and ML methods. The data become more complex if the communication functions support the exchange of messages and *offers* [10].

We work with data collected by the NSS Inspire [9] since 1996, the largest available collection of this kind. We analyze here 2557 recorded negotiations. The negotiations have the following characteristics.

- The problem is the purchase of bicycle parts, with four negotiation issues – price, delivery time, payment time, and return policy, each with several possible values.
- There are two participants: a seller (Itex Manufacturing) and a buyer (Cypress Cycles); every negotiator will participate in only one negotiation.
- Upon logging in, the negotiators fill a pre-negotiation questionnaire with negotiation preferences and personal and background data.
- The negotiators exchange formal offers (tables with numbers from a small fixed set), and possibly messages that either accompany offers or are exchanged between offers.
- A negotiation is completed only if an offer has been accepted and acceptance registered by Inspire within three weeks; it is uncompleted otherwise.

Inspire offers very highly developed, diverse e-negotiation support [11]. It includes: preference assessment in the pre-negotiation phase; offer and message exchange medium, analysis of alternative offers, counter-offer evaluation, access to the on-line manuals and history of the negotiation in the negotiation phase; assessment of the efficiency of the compromise (Pareto-optimality) in the post-settlement phase. Inspire's main decision-analytical tool is the utility function [11], calculated for each negotiator considering the preferences for each value of each negotiation issue and balancing preferences for single-issue values with combinations of values. The user can change the utility function during negotiation, so it is not an objective measure of the negotiation process or its outcome. Few participants answer the post-questionnaires, making the data therein unreliable for generalization. In the end, the Inspire data useful for our studies consist of the text data, the offers, the history records, and the pre-negotiation questionnaires. At present, we do not include the offer data in the representation.

3 Data pre-processing

The 2557 negotiation records collected with Inspire contain approximately 1,514,600 word tokens and 27,000 word types, contributed by more than 5000 authors.

Table 2 shows how much variety there is in the Inspire negotiators' background. 3125 negotiators identified their first language, 4276 their occupation.

First language	%	Occupation	%
English	28.1	students	82.8
German	22.8	professionals	13.1
Mandarin, Kantonese	12.1	managers	1.8
Spanish	9.7	engineers	1.1
Hindu	4.6	teachers	0.6
Russian	3.8	professors	0.4
Finnish	3.4	executives	0.2
Others	15.5		

Table 2: Background of Inspire negotiators

Despite the variation in negotiators' cultural and educational background, the messages they exchanged share some interesting characteristics: they are dense, subject-oriented, points of discussion are often accompanied only by salutations and closure, casual talk appears later in negotiation. In casual talk senders exchange personal information, so it contains geographical names, names of celebrities, names of sport teams, and so on.

English was suggested as the language of negotiation. The level of proficiency in English of the negotiators varies, and occasionally the negotiation was conducted in a different language – French, German, Spanish or Russian transliterated in Latin alphabet. Figure 2 shows a fragment of a dialog extracted from a completed negotiation.

These messages are unedited, and contain much noise. We have identified through manual analysis the following types of noise:

1. Messages with words containing non-letter characters.
2. Text segments in foreign languages, written in ASCII code.
3. Use of foreign words within the English text.
4. Use of informal and slang expressions.
5. Spelling errors, missing punctuation and spaces between words, incorrect capitalization.

We call words affected by noise *noise-corrupted* words. Some of the most frequently misspelled words are **deliver**, **negotiate** and **receive** and their derivatives, and **sincerely** and **unfortunately**.

(**Seller**) *Hi Anles, I have just sent a counter-offer to you. It wasnt such easy, as I thougt cause it seemed I made my ratings wrong *g*. Well, now I already asked you, where you are from, cause I did not know that I would have the opportunity to contact you again. I am from Germany. Then, good luck with my offer, I am waiting for your answer. Bye Claudi*

(**Buyer**) *hi claudi, thank you very much for your offer. I think, the price is acceptable. I totally agree with you. Having informed at a trade fair in Frankfurt/Germany about metal components and comparing some prices and offers from other suppliers all around the world, I came to the conclusion that your offer is the best. It was a pleasure doing business with you. I'll give you a ring this week for more details. Best regards anles (wir wren jetzt wohl schon am ende unserer negotiation. leider war es nicht lang, da ich schon jetzt eine ziemlich hohe punktzahl erreicht habe. du vielleicht auch... ich komme brigens auch aus deutschland. lustig, oder? woher kommst du denn genau? wenn dir das geschreibe ber interneg zu langwierig ist, kannst du mich ja auch per mail erreichen: [...] ok. wrde mich freuen. cu anles (Anja)*

Figure 2: Example message exchange in a negotiation using Inspire

We have observed that different types of noise appear in different places in the data. While noise of type 1 and 2 is concentrated in big chunks, noise of type 3, 4 and 5 is spread throughout the data. Out of these 5 types of noise, only noise of type 1 can be eliminated automatically, the other require manual intervention.

To perform statistical analysis and modelling of the language of e-negotiations, we automatically filter out from the original data set all messages written in languages other than English. Figure 3 shows schematically the filtering process that gives us the data set we work with.

In preliminary studies [17] performed on a subset of the Inspire data, we have proposed that the Inspire vocabulary grows as the vocabulary of unrestricted languages [13]. To test this hypothesis we have computed the growth rates of the type-token ratio ($TT(N)$)[1], of the vocabulary

$$P(N) = \frac{V(1,N)}{N} \tag{1}$$

[1] N is the number of tokens

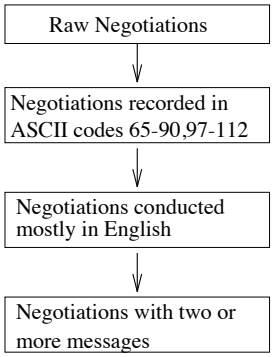

Figure 3: Data preprocessing

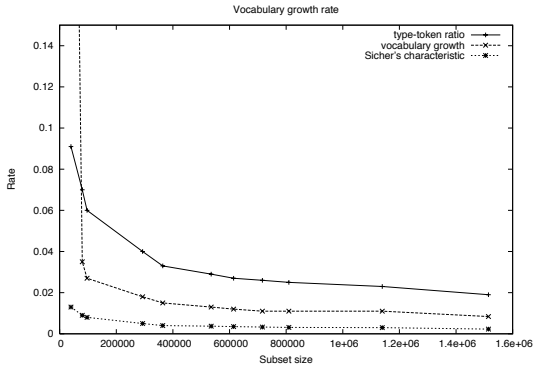

Figure 4: Vocabulary growth rate for Inspire data

and Sichel's characteristic

$$S(N) = \frac{V(2,N)}{N} \tag{2}$$

($V(i,N)$ is the number of types that occur i times in the text with N tokens.)

We show the results in Figure 4. The fact that Sichel's characteristic converges, and that with each increase in the data set the number of rare words increases, proves our hypothesis.

4 Data Modeling

The *Inspire* data show many characteristics interesting for statistical language modeling [14], quite different than the characteristics of such widely used NLP corpora as *Brown* [4] or the *Wall Street Journal* corpus. Unlike those standard corpora, ours has a higher token-type ratio, regular percentage of rare words [7] and high percentage of most frequent words. There also seem to exist no references to work on modeling text data from bilateral CMC. We present here a preliminary study of statistical modeling of such data. N-gram models are arguably the most widely used. An important concern in such modeling is *smoothing* that

helps incorporate the knowledge of the previously unseen N-grams. We use Katz smoothing [8] that generally performs well on data with a high token-type ratio, is easier to implement, has very few parameters and does not require a validation set [1, 7] for adjustable parameters. In our future work, however, we want to incorporate some data-dependent information too, for example, the negotiation-specific distinction between buyers and sellers. The size of our data is another argument for Katz smoothing.

We now briefly describe Katz smoothing and show the cross-entropy results for it and for another attractive technique, Kneser-Ney smoothing. The latter performs well with respect to collocations; this might be helpful since we work in a closed domain with a fixed topic. Katz back-off smoothing [12] basically combines the higher-order models with lower-order models, extending the classical Good-Turing estimate [3].

As usual, we are looking how to estimate $p(w_i|w_{i-1})$, the frequency of appearence of the word w_i given that the last word was w_{i-1}. For simplicity, we consider Katz estimation of $p_{katz}(w_i|w_{i-1})$, which corresponds to the case of a *bigram* model. It will be defined shortly. Katz smoothing for n-gram models of higher orders is analogous. In fact, any Katz n-gram model is defined in terms of the Katz $(n-1)$-gram model.

A sentence s is composed of words $w_1...w_l$. The corrected count of a bigram $w_{i-1}^i = w_{i-1}w_i$ with count $r = c(w_{i-1}^i)$ is:

$$c_{katz}(w_{i-1}^i) = \begin{cases} d_r r & if r > 0 \\ \alpha(w_{i-1}) p_{ML}(w_i) & if r = 0 \end{cases} \tag{3}$$

d_r is the Good-Turing discount ratio which reduces non-zero counts in order to keep the probability of the whole event to one when non-existent (zero) counts are added. $\alpha(w_{i-1})$ is a normalizing parameter which allows to distribute only the probability mass left over in the discounting process. $p_{ML}(w_i)$ is the maximum likelihood estimate of the probability p_{w_i} [12].

We discount all the bigrams with non-zero count with d_r approximately equal to $\frac{r^*}{r}$ where r^* is the Good-Turing estimate of r:

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \tag{4}$$

n_r is the number of N-grams that occur exactly r times in the training data. The counts thus subtracted are distributed among the zero-count bigrams according to the unigram model distribution, that is, the next

lower-order distribution. The value of $\alpha(w_{i-1})$ is chosen with the constraint $\Sigma_{w_i} c_{katz}(w_{i-1}^i) = \Sigma_{w_i} c(w_{i-1}^i)$.

The value of $\alpha(w_{i-1})$ is:

$$\alpha(w_{i-1}) = \frac{1 - \Sigma_{w_i:c(w_{i-1}^i)>0} p_{katz}(w_i|w_{i-1})}{\Sigma_{w_i:c(w_{i-1}^i)=0} p_{ML}(w_i)} = \frac{1 - \Sigma_{w_i:c(w_{i-1}^i)>0} p_{katz}(w_i|w_{i-1})}{1 - \Sigma_{w_i:c(w_{i-1}^i)>0} p_{ML}(w_i)} \tag{5}$$

Now, normalizing yields the value for $p_{katz}(w_i|w_{i-1})$:

$$p_{katz}(w_i|w_{i-1}) = \frac{c_{katz}(w_{i-1}^i)}{\Sigma_{w_i} c_{katz}(w_{i-1}^i)} \tag{6}$$

Large counts are generally considered reliable and hence d_r is taken as 1 for any count $r > k$ where k is suggested by Katz to be 5. For lower counts $r \leq k$, d_r is found according to the following equation:

$$d_r = \frac{\frac{r_*}{r} - \frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}} \tag{7}$$

Finally, the Katz unigram model is taken to be the Maximum Likelihood probability, as we said above. For more details on Katz smoothing see [8, 1].

The practical attractiveness of the Katz model is noteworthy. It is easy to implement, which we did, and it is easy to implement with the additional search for the value of Katz coefficient which gives the smallest value of the cross-entropy, which we also did.

Now, we need to evaluate the goodness of the fit for this model on our data. The standard measure for evaluating statistical models is cross-entropy:

$$-\frac{1}{n} \sum_{i=0}^{n} log(P(w_i)) \tag{8}$$

n is the number of words in the test set, $P(w_i)$ is the probability of the appearance of the word w_i in the test data. A model with the lower cross-entropy on the test set models the data better [1]. With respect to the language, the cross-entropy is one of the estimators of complexity, or predictability of that language [1, 14]. The lower cross-entropy means that the data is predictable, higher cross-entropy indicates high uncertainty in the data. Incidentally, the cross-entropy of English texts ranges from around 5.64 to 9.70, depending on the type of text [1].

The analysis of the applicability of statistical models to the different types of data of the same size as the Inspire data suggest the following:

Table 3: Cross-entropy results

Data	GTK model	KN model	ELE model
partial	5.69	5.75	6.03
whole	5.66	N/A	N/A

(1) trigram models with the modified Kneser-Ney smoothing method (KN) [1] and the Katz variant of Good-Turing smoothing method, with $k = 5$ (GTK), where k is the number of occurrences of a unigram in the data [1]; (2) the Expected Likelihood Estimation (ELE) model [12] that we consider a Baseline model.

In the first step of building the models we used part of the Inspire data, with 648,931 tokens of which 581,631 were in the training set. In the second step we used all Inspire data, 1,107,447 tokens for training and 398,703 for testing (other tokens were filtered out as non-English words). On the partial data, the cross-entropy values obtained by the ELE and KN models were higher than the cross-entropy obtained by the GTK model. The KN model ran significantly longer than the GTK model. We therefore built only the GTK model on the whole data. The cross-entropy reduction of more than 0.160 is noteworthy [14, 1]. In the case of e-negotiation data the reduction was 0.34 compared with the baseline. Such reduction usually corresponds to an improvement of application performance [14]. The cross-entropy results are reported in Table 3.

5 Lexicon Construction

For a deeper and more semantic analysis of the Inspire collection, we need to analyse the vocabulary used by negotiators. We have developed a procedure for extracting a corpus-based lexicon: a monolingual lexicon with general syntactic and domain-specific semantic information.

Misspelled words that cannot be restored are not included in the lexicon. Person names, location names and e-mail addresses are easily identified using the questionaire files and the salutations in the messages exchanged. The rest of the words are spell-checked using an off-the-shelf spell-checker (`ispell`). Frequency counts are used to select the most probable correct spelling for each misspelled word from the options that `ispell` gives. After correction, the words are lemmatized. We use Longman Dictionary of Contemporary English (LDOCE) to extract the subset of lemmas that also appear in the dictionary. We call these

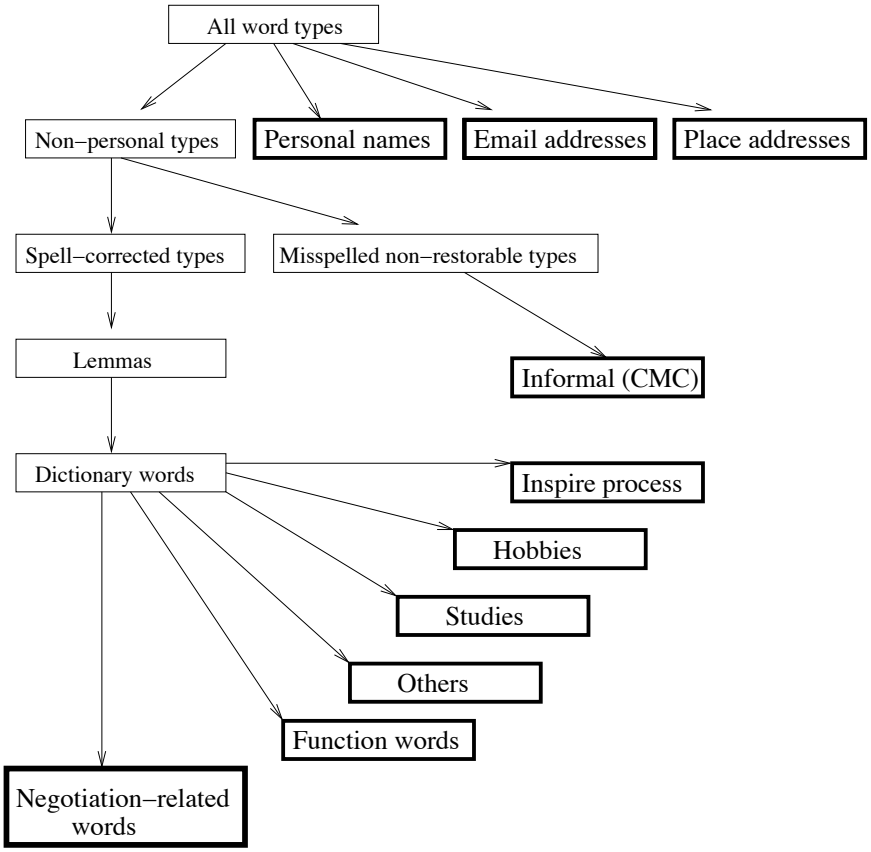

Figure 5: Building a semantic lexicon

dictionary words.

Figure 5 shows the types of words in our data, at various stages in the lexicon-construction procedure.

We would like this lexicon to include syntactic and semantic information about each word. We assigned each word part of speech information using information extracted from LDOCE.

5.1 Adding semantic information

Textual data can be analyzed at different levels of granularity. Throughout our statistical analysis, we considered the basic unit to be a word.

We consider two issues regarding the semantics of the words in our collection: (1) decide on the sense of each word in the particular context

of (electronic) negotiations, (2) group word senses in clusters that would allow us to have a higher-level view of the data.

To deal with the first issue we could take an off-the-shelf dictionary, and assign each word a sense from the ones listed. The problem is that lexical resources are limited, and do not include all possible senses. The word *margin* for example, appears in LDOCE with a sense related to PUBLISHING. In our collection, this word appears in expressions such as *profit margin*, *low margin*, which give it a business-related meaning.

We decided to combine the two issues and address them together. We analyze our data manually and devise a set of semantic clusters, or categories, that cover all the words. We then assign each word to one of these categories.

First of all, the general domain of this data is negotiations. Frequency analysis of unigrams shows that words related to the negotiation process rank high. Our first semantic category is then **negotiation-related words**.

Because the data is collected through Inspire, a Web-based NSS, we find that the users discuss the system itself and its performance. There are also discussions related to CMC. Based on this observations we propose two more semantic categories: **Inspire process** and **Informal (CMC) words.**

Negotiation partners sometimes exchange information not related to the main topic. They exchange contact (names, places, e-mail addresses), personal (hobbies) and professional information. Because 82.8% of negotiators are students, professional information refers mostly to studies. Following these observations we extend the set of semantic categories with **studies**, **hobbies**, **personal names**, **place names**, **e-mail addresses**.

Function words, such as prepositions, are quite frequent, and the category of **function words** groups them all. The last category, **others**, catches the words that do not fit in any of the previous clusters.

Despite the fact that not all word senses appeared in LDOCE, we attempted to find a map between our semantic categories and the categories that LDOCE provides. This allows us to reduce the amount of manual tagging. Table 4 shows the mapping with the smallest number of misclassified examples[2]. From some LDOCE categories we first remove words in subcategories. For example, we remove biological terms (they

[2]Informal CMC words, personal and place names, e-mail addresses do not appear in LDOCE, so no mapping is shown for these categories. Function words are a closed class, no semantic analysis is required for them.

Table 4: Correspondence between the Inspire semantic categories and LDOCE categories.

Data-dependent tag	**LDOCE category tags**
Negotiation-related	Business without daily life, crime and law without daily life, birth, death, publishing bicycles, cars (all of them without biology)
Inspire process	Data processing and computing
Hobbies	General sports, Leisure
Studies	Education (without biology)
Others	Daily life without sports, general society, politics, religion General transport, general engineering, general industry, (all of them without biology)

appear as nicknames in the Inspire data) from the education category, and we mark the remaining words Studies.

We will use the lexicon we have built, and especially the semantic information, to analyze the Inspire data on a more semantic level. Semantic categories will allow us to generalize and look for linguistic patterns that signal various types of behaviour and intentions of the negotiators.

6 Conclusion

We have discussed and presented two methodologies for the analysis of textual data from e-negotiations. We have presented the results of applying them to the Inspire data. The statistical and lexical models we have built reveal specific characteristics of the Inspire data, compared to other corpora used in NLP: noise caused by specific phenomena, vocabulary growth as unrestricted language, specific semantic categories for the words in the data. The results reported cover preliminary research, part of an ongoing project.

More detailed analysis is currently under way. We split the data into subsets of interest, based on the outcome of the negotiations or on the role of the negotiators. We will look for language patterns indicative

of negotiation strategies, and we will explore the interaction between negotiators with various negotiation tactics.

The methodology we are building is not particular to the Inspire collection. It can be applied to any CMC data that share the following characteristics: they are collected from goal-oriented communications in which participants have well-defined roles, the purpose of communication is specified and there are clear criteria that define possible outcomes, duration, obligations, and acceptable behaviour (for example, the message exchange between a doctor and a patient, a counsellor and a student, and so on).

With the increase of Web-based business communications, the need for such a methodology increases.

Acknowledgment

Support for this work includes a major grant from the Social Sciences and Humanities Research Council of Canada and a doctoral scholarship from the Natural Sciences and Engineering Research Council of Canada.

References

[1] Chen S. F. and J. Goodman (1998), *An Empirical Study of Smoothing Techniques for Language Modeling*, Harvard University, Center for Research in Computing Technology TR-10-98.

[2] Climent S., J. Moré, A.Oliver, M.Salvatierra, I.Sánchez, M. Taulé and L. Vallmanya (2003), *Bilingual Newsgroups in Catalonia: A Challenge for Machine Translation*, J. Computer-Mediated Communication, **9(1)**, [http://www.ascusc.org/jcmc/vol9/issue1/climent.html].

[3] Good I. J. (1953), *The population frequencies of species and estimation of population parameters*, Biometrika, **40(3,4)**, 237–269.

[4] W. N. Francis and H. Kucera(1979), *Brown Corpus Manual*,**http://helmer.aksis.uib.no/icame/brown/bcm.html**.

[5] Hargie O. and D. Dickson (2004), *Skilled Interpersonal Communication: Research, Theory and Practice*, 4th ed., Routledge.

[6] Herring S. C. (2001, *Computer-mediated discourse*, In D. Tannen, D. Schiffin and H. Hamilton (eds.), *Handbook of discourse analysis*, Blackwell, 612–634.

[7] Jurafsky D. and J. H. Martin (2000), *Speech and Language Processing*, Prentice Hall.

[8] Katz S. M. (1987), *Estimation of probabilities from sparse data for the language model component of a speech recognizer*, IEEE Transactions on Acoustics, Speech and Signal Processing, **35(3)**, 400–401.

[9] Kersten G. E. *et al.* (2002-2006), *Electronic negotiations, media and transactions for socio-economic interactions*, [http://interneg.org/enegotiation/].

[10] Kersten G. E. (2000), *Modeling Distribution and Integrative Negotiations. Review and Revised Characterization*, Group Decision and Negotiation, **10(6)**, 493–514.

[11] Kersten G. E. and S. J. Noronha (1999), *WWW-based Negotiation Support: Design, Implementation, and Use*, Decision Support Systems, **25**, 135–154.

[12] Manning C. D. and H. Schutze (1999), *Foundations of Statistical Natural Language Processing*, 4th ed., MIT Press.

[13] McEnery T. and A. Wilson (2001), *Corpus Linguistics*, Edinburg University Press.

[14] Rosenfeld R. (2000), *Two Decades of Statistical Language Modeling: Where Do We Go from There?*, Proc IEEE, **88**, 1270–1278.

[15] Schoop M. (2003), *A Language-Action Approach to Electronic Negotiations*, Proc 8th International Working Conference on the Language-Action Perspective on Communication Modelling (LAP 2003), 143–160.

[16] Shah M., M. Sokolova and S. Szpakowicz (2004), *Using Domain-Specific Knowledge to Classify E-negotiations*, InterNeg Report 07/04, [http://interneg.org/interneg/research/papers/].

[17] Sokolova M., S. Szpakowicz and V. Nastase (2004), *Using language to Determine Success in Negotiations: A Preliminary Study*, Proc

17th Conference of the Canadian Society for Computational Studies of Intelligence, 449–453.

[18] Thompson L. and J. Nadler (2002), *Negotiating Via Information Technology: Theory and Application*, Journal of Social Issues, **58(1)**, 109–124.