STRUCTURED AND QUANTITATIVE PROPERTIES OF ARABIC SMS-BASED CLASSIFIED ADS SUBLANGUAGES

Daoud M. Daoud[1], Samir A. El-Seoud[2] and Christian Boitet[3]

[1]PSUT, King Hussein Faculty of Computing Sciences - Jordan
[2]Faculty of Informatics and Computer Science, The British University in Egypt – BUE
[3]à l'Université Joseph Fourier (Grenoble 1), UFR IMA

ABSTRACT

In this paper we will present our work in studying the sublanguage of Arabic SMS-based classified ads. This study is presented from the developer's point of view. We will use the corpus collected from an operational system, CATS. We also compare the SMS-based and the Web-based messages. We also discuss some quantitative properties of the studied text.

KEYWORDS

Arabic Classified Ads, Sublanguages, SMS-based corpus

1. INTRODUCTION

In order to better understand the sublanguage used in SMS-based classified ads and similar applications; the approach of corpus-based is certainly needed. Moreover, it helps to understand how people encode their thoughts in the context of the device, task and domain.

In turn, this will help in selecting the right approach for the development of NL systems. As an example, systems developed for semi-structured text are not appropriate for free text and vice-versa. Keeping in mind that text editing requires information systems different than those developed for natural, spontaneous and unprocessed text.

Most of the current systems that process users' queries and generate responses use shallow text processing techniques based on pattern extraction or information retrieval techniques [2].

Systems that simplifying the user interaction styles by using either form filling or controlled language [7] and [8] are doing so in order to avoid the hard problem of supporting free natural language interface. Furthermore, other systems that try to use inappropriate NLP techniques to support free NL interface failed.

Understand the way in which people encode their thoughts plus represent the model of the concerned domain knowledge correctly are the success factor to build NL systems related task.
The shortage of data is one of the main obstacles in developing natural language systems. It is not easy to collect corpuses for restricted domains, especially if they must come from a very private medium of communication such as SMS. A very interesting aspect of CATS [6] and [8] is the study of its natural "sublanguages". That provides an unprecedented opportunity to analyze the SMS-based restricted domain sublanguages As a matter of fact, the dataset collected from CATS

DOI: 10.5121/ijnlc.2016.5101

is unique and very interesting. Firstly, it contains real, spontaneous, and unedited text. Secondly, it is written by thousands of authors from a diversity of backgrounds. Thirdly, it covers different domains and topics.

The experiments conducted by [17] clearly concluded that parsing a domain-specific text requires a grammar that suits that domain. Domain-specific grammars can produce better results than domain-independent grammars. From a practical perspective, corpuses are the sources of these domain-dependent grammars.

We will first discuss related works. Then, we will analyze the typology of SMS-based sublanguages. Firstly, we will present some quantitative properties. Finally, we will present the lexical, syntactic and semantic features of these types of sublanguages.

2. RELATED WORKS

It is noticeable that in restricted domains of knowledge, among certain groups of people and in particular types of texts, people have their own way of encoding their thoughts. Such restrictions reduce the degree of syntactic and lexical variation in text[16].

Examples of sublanguages are the languages of weather bulletins, aircraft maintenance manuals, scientific articles about pharmacology, hospital radiology reports, and real estate advertisements [12].

The study of a sublanguage corpus is necessary for producing a sublanguage grammar. As presented in figure 1, the analysis of the linguistic aspects and features of a sublanguage is needed to specify the sublanguage grammar (with the incorporation of the domain knowledge). Therefore, general linguistic knowledge and sublanguage grammar can be used to determine the best NL technique to use. Similarly, the sublanguage grammar and the domain knowledge are both indispensable in selecting the best content representation.

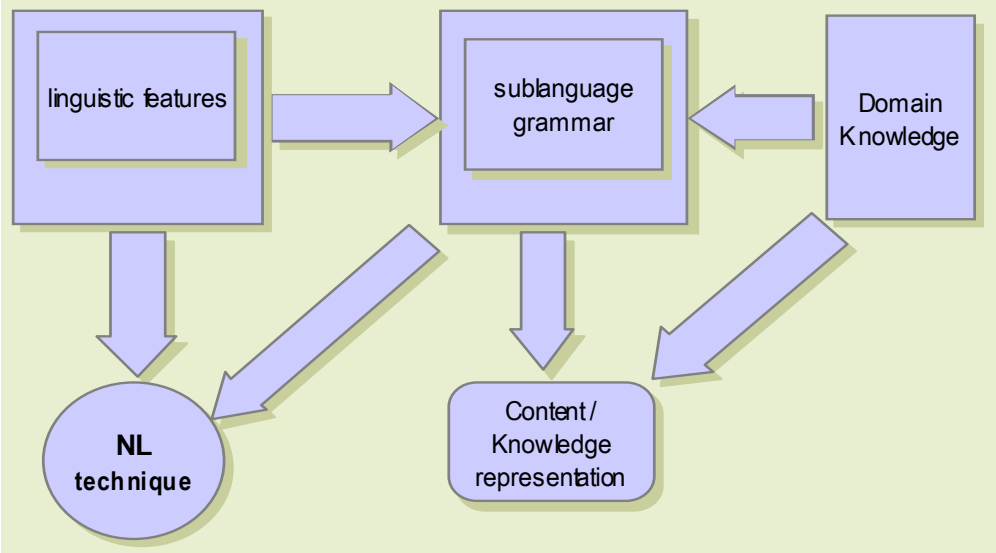

Figure 1. NL development using sublanguage study

Most researchers studied sublanguages from the regularities within syntactic features [3]. The author in [18] uses the type-token ratio (TTR) in the analysis for different corpuses. She found that highly technical writing has a lower TTR than technical writing. [14] Studied four technical

sublanguages (weather bulletins, recipes, stock market reports, and aviation maintenance manuals) and found syntactic homogeneity within each sublanguage. [3] compared the sublanguages of hard sciences and social sciences in their use of various syntactic patterns.[4], studying classified ads, used the type-token ratio (TTR) in his analysis of different levels of text. He found that the jobs domain has a higher ratio than the cars and apartments domains. [10] Studied a clinical sublanguage and a second sublanguage, biomolecular literature. Both of them studied by establishing semantic categories for the entities and relations in the domain, specifying semantic and syntactic co-occurrence patterns, and specifying target forms for each of the patterns. [19] Studied a German-language corpus that contains different medical texts using TTR and sentence length.

3. TYPOLOGY OF SMS-BASED TASK-ORIENTED SUBLANGUAGES

The type-token ratio (TTR) has been used for the measurement of the lexical complexity of SMS-based classified ads sublanguage. Richness of text as well as lexical complexity causes the TTR to be increased. Otherwise less lexical complexity and less repetition of the same word cause TTR to be decreased. For fair comparisons, the authors in this paper use different corpuses to calculate TTR.

We measure the language complexity by the length of the sentence in words, and we do comparisons between different domain and between SMS-based and Web-based sentences.
Finally, the nature of the text, i.e. telegraphic or normal, is identified by the frequency of the words in a corpus. The less function words percentage in a corpus, the more break into pieces in the text style increasing

The manual study of lexico-semantic patterns found in the posts is a part of the analysis of the sublanguage. Our objective is extracting classes of objects that specify the domain knowledge described by the sublanguage.

3.1. GENERAL CORPUS STATISTICS

Within a limited period of CATS operation, a collection of posts from Cars and Real Estate domains constitutes the SMS-based corpus (see table 1 below). We also collected SMS-based Job announcements sent to another mobile short number connected to a publisher of a printed circulation interested in the Jobs domain. The open domain sentences are the posts received by CATS or by the Job announcement short number and not related to any mentioned domains or any related ones.

Table 1. Examined SMS-based Corpus

Domain	Number of sentences	Sentence average length (words)	types	Tokens	TTR
Cars	771	9	1181	5875	.201
Real Estate	641	12.5	1441	6182	.233
Jobs	174	14.61	921	2452	.375
Open	231	6.42	1099	1538	.714

Compared to 7.3 words for TREC questions, the Real Estate domain length of sentences is bigger than the Cars domain length of sentences, see figure 2. We also find that the average sentence length is highest in the Job domain. These findings suggest that the language complexity, as measured by sentence length, is higher in the Jobs domain. It means more amounts of words are needed to encode thoughts in the Cars domain than in the in the Jobs domain or in the Real Estate domain.

We also find that the average length of SMS-based open domain sentences is only 6.42 words, making it the smallest among all other domains. SMS-based posts are generally smaller compared with Web-based posts. This might be explained by the fact that the SMS medium imposes a length and cost constraint on the advertisement. On the other hand, the Web posters have no strict constraint, are encouraged to making their posts arbitrarily long, and therefore tend to include more irrelevant information.

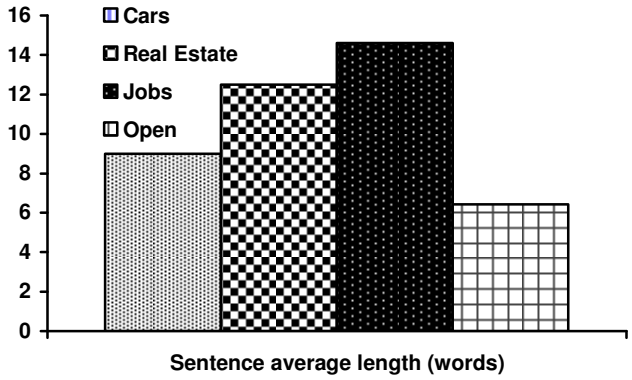

Figure 2. Average message length for each domain

We also calculated the type-token ratio (TTR), the ratio of the number of word types to the number of word tokens in the text, as a measure of range of vocabulary used. That is, the greater the range of possible referents, the greater the number of word types which will appear in a text. For a given number of tokens, a text tending toward greater explicitness will contain a higher number of types [4].

As shown in figure 3 and figure 4, the least TTR value was for Cars at 0.201, then for Real Estate at 0.233, then for Jobs at 0.375. That suggests that the vocabulary is more limited in the Cars domain than in the Real estate domain, and in turn than in Jobs announcement.

We also find that the TTR of the SMS-open domain is very high (0.714), suggesting a lexically complex and rich text. This indicates that posters were not focusing narrowly on particular topics of discussion so that the same words were not repeated often. These TTR results confirm the findings of [4] for the same domains.

Compared with SMS based posts, the values of the type-token ratio for Web-based posts were found to be lower. Therefore, a higher lexical complexity and diversity in the SMS-based text is highly recommended.

The type-token ratio of general Arabic corpus of nearly the same number of token (i.e. same text length) is 0.539 as it has been computed in [11].

Comparing this result with TTRs found for the Cars, Real estate, and Jobs posts confirms the narrow scope and limited vocabulary characteristic of SMS-based restricted domains. On the other hand, the SMS-based open domain TTR is higher than that of [11], suggesting a more topical diversity, see table 2.

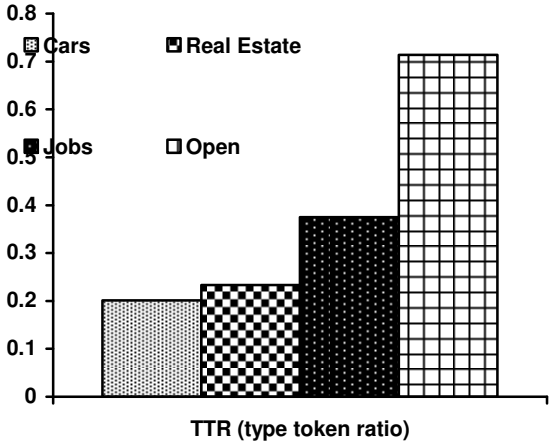

Figure 3. TTR of SMS based messages

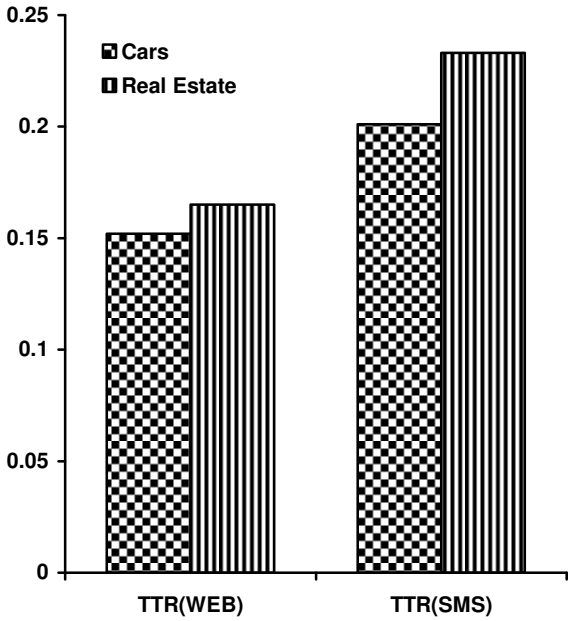

Figure 4. Web-based compared to SMS-based TTRs

Table 2. Comparison of SMS based TTRs with the findings in [11]

Domain	Types	Tokens	TTR	TTR of general Arabic
Cars	1181	5875	.201	0.539
Real Estate	1441	6182	.233	0.539
Jobs	921	2452	.375	0.579
open	1099	1538	.714	0.618

Figure 5 measures the new types added for each 200 messages in the Cars and Real Estate domains. We find that the Cars new types are less numerous than in Real Estate. We also observe that the difference of new types for both domains is proportional to the number of messages. This suggests that the Real Estate domain is richer in vocabulary than the Cars domain, a finding that does not contradict the TTR findings.

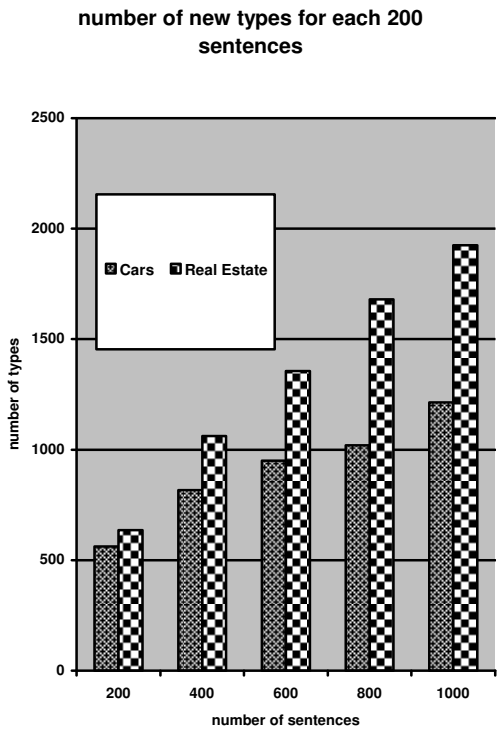

Figure 5. Number for new types for each 200 messages for the Cars and Real Estate Domains

3.2. WORD FREQUENCIES

Cars 53.77%, Real Estate 45.76%, Jobs 43.27% and open domains 22.37% are the top 50 most frequently used percentage of words in SMS-based.

These findings suggest that as we move from Cars to Real Estate, to Jobs and finally to open domain, the percentage of function words (such as prepositions) increases. This finding can be correlated with the TTR of each sub-domain, indicating a less telegraphic text as we move from the Cars domain to the open domain.

3.3. LEXICAL CHARACTERISTICS

The low TTR values of SMS-based Cars or Real Estate domains compared to those of the open domain confirm the fact that the most obvious feature of a classified ads sublanguage is its limited and specialized lexicon. Although the limited vocabulary same concept in posters might be expressed using different words. For example, posters might use about 30 words and spelling variations for the concept "more".

3.3.1. TERMINOLOGY

For some domains, such as Cars domains as well as Real Estate domains, some words might have different meanings than in the open domain. For example, *'duck'* refers to a Mercedes model in the Cars domain and *'piece'* refers to a land in the Real Estate domain. For this reason, to process the text a specialized dictionary is required. Model and a *'piece'* mean a land.

At the topmost frequent list of words, Multi-word concepts and terms appear regularly. As an example:

"For sale"
"For rent"
"Looking for"
"Piece of land"
"Not more than"
"Full option"
"Except the transmission"

They take a special meaning as if they were single words. Whether to treat them as a single words or not is an issue debated among researchers. However, since they are very discriminating in restricted domains, handling them as single words concepts is very useful to improve the precision of such systems [9].

3.3.2. ABBREVIATIONS

The use of abbreviations is not frequent, and there is no common standard or agreement among posters. As an example "م" denotes *area*, *meter* (Real estate domain) or *model* (Cars domain). In the same context, most posters do not abbreviate "West Amman". However when some of them do, we get several abbreviations: "عمان غ", " ع غ", " غ ع" , " غ عمان".

3.3.3. NAMED ENTITIES

Named entities in the car domain refer to model and car makes. While named entities in the Real Estate domain refer to Locations. Named entities in the Open domain might be extended to include dates, number expressions and times, dollar amounts as well as web and email addresses. Correct recognition of named entities is necessary in raising the quality of the processing. As an example, it is impossible to process "For sale Honda" properly without the recognition of "Honda" as a Car Make.

The research studies of Arabic SMS-Based classified ads corpus show that Named Entities could consist of one or more words. We also note that it is possible that each component (or combinations of them) of a multi-word named entity points to a different reference or has another meaning. For example, the multi-word named entity "Biader Wadi AlSir" is a location name referencing a well known area in Amman, but "Wadi AlSir" alone references other area. As Arabic is not like English in distinguishing named entities by capitalizing the first character, and sentences are very short, recognition of named entities is impossible without using lexical lookup.

3.3.4. NUMERICAL VALUES

The dataset under study is full of numerical values. Numerical values that represent year, price, model, and motor size exist in car domain. While numerical values that represent number of bedrooms, area, price,…etc exist in the Real Estate domain. The posters encode numerical values differently. Some of them use non-Arabic numerals such as "three thousands". Others use Arabic numerals such as "3000". Moreover, an expressions such as "3 thousands" might be used in some posters as a combination of the two approaches.

Usually, numerical values are preceded by hint words such as:

Area
Price
Motor size
year

Some of them are usually followed by unit words such as:
Meter
CC
Dinar

Both hint words and unit words are used by the rules to extract correctly the values. Users should write both hint words and unit words as demonstrated by the post. Otherwise, it might cause a problem. For example:

“For sale Mercedes 200 1999”

Failing to write hint words or unit words near to numerical values is frequent within the studied corpuses.

3.3.5. SLANG WORDS

The corpus also contains slang words and words from spoken Arabic like "بدى" “I want”. They are not so frequent (5 slang words per 100 SMS messages) in the SMS-based classified ads domain as none of these slang words appears in the top 50 most frequent words. Nevertheless, they need to be processed accurately.

3.3.6. SPELLING SYSTEM

Many variations for Arabic text spelling may exist in the studied corpus. It is very important to analyze them from the development point of view.

For example, people write the Alef letter “ا”, or with Hamza (ء) over it “أ” or under it “إ”. Also, we find confusions between the Ha’ “ه” and Ta’ “ة”, and between Ya’ “ي” and Alef-Maqsoura “ى”.

The wrong insertion of spaces is another problem. Normally, spaces are used to separate words in Arabic. Most Arabic letters are connected from both sides (cursive writing system), causing them to have different shapes depending on their position (first, middle, or last). But the letters “و”, “ر” ,”ز” , ”د”, ”ى” and ”ذ” can be connected only from the right side, making their shapes unchanging at any position within a word. People tend to wrongly insert or omit a space after any of these letters: : (e.g. Abu-Baker: "أبو بكـر" or "أبوبكـــــر")

The inconsistency of the Arabic spelling of transliterated proper nouns is also detected in the classified ads text where many of the proper names (car make and model as an example) are transliterated from other languages. This phenomenon is noticeable within unedited and spontaneous classified ads, reflecting the cultural and educational background of the text writers. For example, the car-make CITROEN has different spellings in our corpus:

{SATARWEN} “ســـترون”
{SA:TERWEN} “ســـاترون”
{SATERWE:N} “ســـتروين”
{SA:TERWE:N} “ســـاتروين”
{SE:TERWE:N} “ســـيتروين”

3.4. SYNTACTIC CHARACTERISTICS

Many alternative surface structures for the same utterance are contained in the data we studied. That phenomenon is due the diversity of the posters. Evidently, in the sublanguage used there was no unique underlying syntactic structure, i.e. some posts consist of fragmented telegraphic phrases, other posts are more cohesive while some other posts consist of full sentences.

Obviously, syntax-based parsing based methods would not prove very useful in dealing with the given data. For example, for sentence 3 in table 3 will not be analysed correctly, if a traditional parser is looking for object and subject. Similarly, the same failure will take place in case of techniques used for semi-structured text that is relying on layout, position and text format.
We also observe the variable order of words in the posts, as shown in table 3. This is a demonstration of the extremely free ordered nature of the Arabic language in which the constituents are not identified by their positions but by their inflectional endings [5].

Additionally, the posts can have different syntactic structures caused by different word orders and grouping patterns of their constituents. The variations extend to the constituents level. The positions of some constituents vary from one post to another: “for sale Honda …..” “A Honda car ……. For sale”. Also, in the Real Estate domain, consider the following posts, which all have the same meaning but in different structures. “…..located between villas), “…..surrounded by villas” “……class A residential area”.

In some posts, some constituents that do not interest the poster, or they are irrelevant from his point of view, i.e. such as “looking for a car model after 2010”, are not present, only one criteria is mentioned and all other criteria are omitted. Some other constituents are omitted because they are implicitly known, for example “looking for a Peugeot 406” in which the word “car” is omitted, or “for sale 600 square meters” in which “land” is omitted.

In some posts, we don’t find any indication of the type (“sell” or “looking for”): “a Toyota Corolla above 99 and with less than 7000 dinar” because the poster thinks it can be known from the context of the post.

Table 3. Top most frequent words in SMS concerning the studied open domain and jobs domain

Open domain			Jobs domain		
%		Word	%		word
1.50	From	من	4.77	in	في
1.11	if possible	ممكن	2.94	young man	شاب
0.98	On	على	2.12	for	عن
0.91	To	الى	1.92	job	عمل
0.91	I am	انا	1.88	experience	خبره
0.91	In	في	1.75	from	من
0.72	know	اتعرف	1.51	on	على
0.72	Ya	يا	1.51	phone	هاتف
0.65	On	علي	1.22	graduate	خريج
0.59	I am	أنا	1.14	looking	يبحث
0.59	Abu	ابو	1.10	with a grade	بتقدير
0.59	To	الي	1.06	good	جيد
0.59	Like	حاب	0.98	diploma	دبلوم
0.52	about	عن	0.98	young girl	فتاه
0.46	know	أتعرف	0.94	and	و
0.46	duleimi	الدليمي	0.82	or	او
0.46	Like	حابه	0.82	university	جامعة
0.46	morning	صباح	0.77	graduate(f)	خريجة
0.46	every	كل	0.77	year	سنه
0.39	I would	ارجو	0.77	years	سنوات
0.39	Jordan	الاردن	0.73	the job	العمل
0.39	I want	بدي	0.73	tel	ت
0.39	greeting	تحيه	0.73	holding	حاصل
0.39	congratulations	مبروك	0.73	experience	خبرة
0.33	To	ألى	0.73	field	مجال
0.33	I love you	احبك	0.61	Bachelor	بكالوريوس
0.33	Allah	الله	0.57	young girl	فتاة
0.33	Girl	بنت	0.53	company	شركه
0.33	Mohammad	محمد	0.49	any	اي
0.33	Hi	مرحبا	0.49	specialty	تخصص
0.33	Hi	هلا	0.49	accounting	محاسبه
0.26	thousand	ألف	0.45	looking	تبحث
0.26	sweeter	احلى	0.41	science	علوم
0.26	good	الخير	0.41	engineer	مهندس
0.26	Iraq	العراق	0.41	job	وظيفه
0.26	doing	العمل	0.37	part time	بدوام
0.26	present	اهداء	0.37	bachelor	بكالوريس
0.26	I love you	بحبك	0.37	part	جزئي
0.26	good	بخير	0.37	holding	حاصله
0.26	Hi	تحية	0.33	or	أو
0.26	How are you	حالك	0.33	looking	ابحث
0.26	Hussien	حسين	0.33	university	جامعه
0.26	number	رقم	0.33	computer	حاسوب
0.26	tribes	عشائر	0.33	year	سنة
0.26	who	مين	0.33	company	شركة
0.20	With	مع	0.33	certificate	شهادة
0.20	With you	معكم	0.33	I have	لدى
0.20	if possible	أذا	0.33	looking	يطلب
0.20	To	أن	0.29	icdl	icdl
0.20	Ahmad	احمد	0.29	ready	استعداد
22.37			43.27		**total**

Table 4. Top most frequent words in SMS concerning the studied cars domain and real state domain

Cars domain			Real Estate domain		
%		word	%		word
5.91	wanted	مطلوب	5.87	In	في
5.86	car	سياره	4.69	wanted	مطلوب
4.82	year	موديل	3.15	for sale	للبيع
4.03	for sale	للبيع	2.35	apartment	شقه
3.13	car	سيارة	1.96	land	ارض
1.70	Dinar	دينار	1.57	apartment	شقة
1.45	check	فحص	1.46	Dinar	دينار
1.34	full	كامل	1.29	Amman	عمان
1.31	full	فل	1.12	than	عن
1.29	Mercedes	مرسيدس	1.08	For a price	بسعر
1.21	color	لون	1.04	not	لا
1.02	for a price	بسعر	0.95	for rent	للايجار
0.95	above	فوق	0.94	meter	م
0.94	Kia	كيا	0.92	from	من
0.92	year	م	0.91	or	او
0.77	and	فما	0.86	land	أرض
0.71	the price	السعر	0.82	house	بيت
0.71	in condition	بحاله	0.73	thousand	الف
0.71	Hyundai	هونداي	0.70	villa	فيلا
0.70	options	الاضافات	0.70	bedroom	نوم
0.70	Daewoo	دايو	0.65	on	على
0.68	except	عدا	0.57	more	يزيد
0.65	bus	باص	0.55	donom	دونم
0.63	gear	الجير	0.55	area	مساحة
0.63	buying	شراء	0.53	meter	متر
0.61	Honda	هوندا	0.50	marj	مرج
0.56	or	او	0.49	piece	قطعة
0.54	Opel	اوبل	0.49	area	مساحه
0.51		ما	0.49	with	مع
0.49	good	جيده	0.47	alhamam	الحمام
0.49	payment	دفعه	0.45	al3ali	العلي
0.46	without	بدون	0.42	with area	بمساحة
0.46	B	بي	0.40	the owner	المالك
0.46	Toyota	تويوتا	0.40	street	شارع
0.46	wanted	طلب	0.39	tela3	تلاع
0.44	power	بور	0.39	jabal	جبل
0.44	above	عن	0.37	saloon	صالون
0.44	Nissan	نيسان	0.37	house	منزل
0.43	BMW	bmw	0.36	thousand	ألف
0.41	Golf	جولف	0.36	west	الغربيه
0.39	installments	اقساط	0.36	3arajan	عرجان
0.39	with payment	بدفعه	0.36	near	قرب
0.39	pick up	بكب	0.36	piece	قطعه
0.39	Lancer	لانسر	0.36	area	منطقة
0.39	from	من	0.36	and	و
0.37	Peugeot	بيجو	0.34	or	أو
0.37	not	لا	0.34	the price	السعر
0.37	and	و	0.34	west	الغربية
0.36	or	أو	0.34	floor	طابق
0.34	new	جديد	0.34	shop	محل
53.77			45.76		**total**

3.5. SEMANTIC CHARACTERISTICS

We have shown that the syntactic structure for different posts which express the same information can vary enormously.

Frequently, many posters are encoding the details of their knowledge at different levels. For example: “looking for a BMW 520” or “A German BMW 520 car for sale”.

The generalization use that exists in the query, i.e. generalization concept for searching, is also presented in the studied corpus. For example, “looking for a German car”, “looking for a villa in Alexandria - Egypt” or “looking for economical car” is frequently used for the searching concept. Implicitly known words like “German”, “Egypt” and “economical” usually do not appear in the "sell" post.

3.5.1. SUBLANGUAGE GRAMMAR

Many researchers believe that it is possible to group words into equivalence classes and to describe the occurring sentences within a sublanguage in terms of these classes [15].

Classified ads post could be viewed as sequence of restricted properties. It restricts the main domain object, for example, apartment and car. For Cars and Real Estate and or "looking for" and "sell" posts the previous statement is true. Instead of relying on syntactic structures for the description of the SMS, it will be more efficient to rely on this information model.

For example, sentence number 1 in table 4 can be described as follows:

[vehicle] [make][model][Ads_type][year][feature]
Year = [year_hint] [number]
Likewise, sentence number 6 can be described as:
[Ads_type][vehicle][price][feature][feature]
Price = [not more][number][currency_unit]

It is clear that the above semantic categories depict the underlying subject matter of the domain. Using pure syntactic description [13] has worst results than describing sentences semantically. It allows semantic knowledge to be easily included in the system [1].

Table 5. Sample of posts in the Cars domain

1	سياره اوبل فكترا للبيع موديل 2003 فل اوبشن	Opel Vectra car for sale year2000 full option
2	للبيع سيارة بي ام دبليو 520 لون زيتي فحص كامل م 89 فل عدا الفتحه مرخصه بحال ممتازه بسعر 8500	For sale BMW 520 color dark green full check year 89 full except sunroof licensed in a good condition with a price 8500.
3	أوبل أسترا ستيشن لون أحمر(بورفتحه سنترزجاج ومريات كهرباء) فحص للبيع.	Opel Astra station color red (power sunroof Center Electrical windows and mirrors check for sale
4	اريدبيع سياره دايو ليمنز موديل 92 فحص كامل فل اوبشن	I want to sell a Daewoo Lemens car year 92 full check full option
5	شراء سيارة.	Buying a car
6	بحاجه لسياره لا تزيد عن 2000 دينار بحاله جيده واقتصاديه في البنزين	In need for a car not more than 2000 dinar in good condition and economical in fuel.
7	عندي سياره لاند روفر بدي ابيعها.	I have a Land Rover car I want to sell it
8	مطلوب سيارة بيجو406	Wanted a Peugeot 406 car
9	لاندروفر ديسكفري 2001 فل 8 سلندر بحاله ممتازه للبيع	Landrover Discovery 2001 full 8 cylinders in excellent condition for sale.
10	سياره للبيع بدفعه 500 هونداي مديل 97 و بالتقسيط	A car for sale Hyundai year 97 by installments.

3.5.2. SEMANTIC CLASSIFICATION

Each concept has a semantic category which is the direct outcome of the study of the corpus. For example, in the Cars domain, we have semantic categories for: vehicle type, car manufacturer, model, colour, motor size unit, motor size hint word, price hint word, currency, features etc.
In the same manner, we have different categories for the Real Estate domain such as: property type, area hint, area unit, locations, floor hint, bedroom hint, feature, etc.

The semantic classification includes semantic taxonomy. We use feature-based categorization of concepts, in which concepts are assigned to categories according to commonalities in specific features. For example, Renault, Peugeot and Citroen share the same feature in that they are all French cars, see figure 6. In a similar manner, Clio and Megan are car models that share the same manufacturer. Similarly, in the Real Estate domain, locations names are grouped together according to the larger area they belong to.

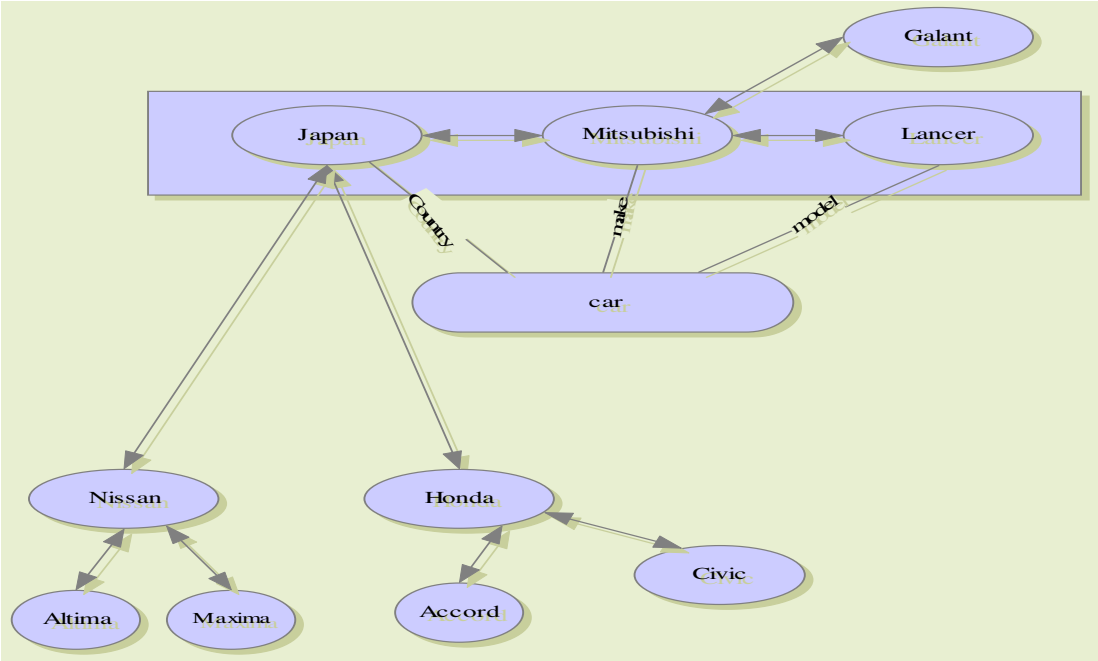

Figure 6. Example of the semantic taxonomy of the Cars domain

4. CONCLUSION

In this paper, we described the structure and the quantitative properties of Arabic SMS-based classified ads sublanguage. From a development point of view, to be able to characterize such restricted languages is quite valuable in that it allows us to detect similar sublanguages and to use the most appropriate robust and effective domain-specific e-commerce language technologies.
In most sublanguages studied, for example, we can show that content-oriented methods are far more efficient than traditional structure-oriented methods, and that is vital to build applications such as content information extraction or e-commerce question request services based on handling spontaneous NL utterances.

REFERENCES

[1] Androutsopoulos I, Ritchie GD, and Thanisch P (1995) "Natural Language Interfaces to Databases: An Introduction", Journal of Natural Language Engineering, Vol. 1, N° 1.

[2] Benamara F (2004) "Cooperative Question Answering in Restricted Domains:the WEBCOOP Experiment", Proc. of ACL'04 Workshop on Question Answering in Restricted Domains, Barcelona.

[3] Bonzi S (1990) "Syntactic Patterns in Scientific Sublanguages: A Study of Four Disciplines", Journal of the American Society for Information Science Vol. 41, N° 2, pp. p121-131.

[4] Bruthiaux P (1994) "Functional variation in the language of classified ads.", Perspectives: Working Papers of the Department of English, City Polytechnic of Hong Kong, Vol. 6, N° 2, pp. 21-40.

[5] Covington MA (1992) A dependency parser for variable–word–order languages, In Computer assisted modeling on the IBM 3090: Papers from the 1989 IBM Supercomputing Competition, edited by Billingsley KR, Brown III HU and Derohanes E, Athens, Greece, Baldwin Press, pp. 799–845.

[6] Daoud D (2005) "Building SMS-based System using Information Extraction Technology", Proc. of ACIDCA-ICMI'2005, Tozeur, Tunisia., 5th to 7th November.

[7] Daoud D (2006) It is necessary and possible to build (multilingual) NL-based restricted e-commerce systems with mixed sublanguage and content-oriented methods, Ph. D. Dissertation, Université Joseph-Fourier-Grenoble I, France, https://tel.archives-ouvertes.fr/tel-00097826/document

[8] Daoud D and Boitet, C (2007) "A STUDY OF ARABIC SMS-BASED CLASSIFIED ADS SUBLANGUAGES", presented at The 1st international conference on digital communications and computer applications (DCCA2007), Jordan

[9] Doan Nguyen H, and Kosseim L (2004) "The Problem of Precision in Restricted-Domain Question Answering – Some Proposed Methods of Improvement", Proc. of ACL'04 Workshop on Question Answering in Restricted Domains, Barcelona.

[10] Friedman C, Kra P, and Rzhetskya A (2002) "Two biomedical sublanguages: a description based on the theories of Zellig Harris", Journal of Biomedical Informatics, Vol. 35, N° 4, pp. 222–235.

[11] Goweder A, and De Roeck A (2001) "Assessment of a significant Arabic corpus", Proc. of Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, July.

[12] Grishman R (2001) "Adaptive Information Extraction and Sublanguage Analysis", Proc. of IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, Washington State, Seattle.

[13] Kate, R. (2012). "Unsupervised grammar induction of clinical report sublanguage." , Journal of Biomedical Semantics, C7 - S4 3(3): 1-13.

[14] Kittredge R (1982) Variation and homogeneity of sublanguage, In Sublanguage: Studies of language in restricted semantic domains, edited by Kittredge R and Lehrberger J, Berlin and New York, de Gruyter, pp. 107--137.

[15] Kittredge RI (1982) "Sublanguages", American Journal of Computational Linguistics, Vol. 8, N° 2, pp. 79-84.

[16] Lehrberger J (1982) Automatic Translation and the Concept of Sublanguage, In Sublanguage: Studies of Language in Restricted Semantic Domains, edited by Kittredge R and Lehrberger J, Berlin & New York, Walter de Gruyter, pp. 81-106.

[17] Sekine S (1997) "The Domain Dependence of Parsing", Proc. of Applied Natural Language Processing(ANLP'97), Washington D.C., USA,, pp. 96-102.

[18] Tagliacozzo R (1976) "Levels of technicality in scientific communication", Information Processing and Management Vol. 12, N° 2, pp. 95-110.

[19] Wermter J, and Hahn U (2004) "An Annotated German-Language Medical Text Corpus as Language Resource", Proc. of 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, 26-27-28 May.