Learning meanings of words and constructions, grounded in a virtual game

Hilke Reckman
The Media Laboratory
MIT
USA
reckman@media.mit.edu

Jeff Orkin
The Media Laboratory
MIT
USA
jorkin@media.mit.edu

Deb Roy
The Media Laboratory
MIT
USA
dkroy@media.mit.edu

Abstract

We discuss the use of data from a virtual world game for automated learning of words and grammatical constructions and their meanings. The language data is an integral part of the social interaction in the game and consists of chat dialogue, which is only constrained by the cultural context, as set by the nature of the provided virtual environment. This paper presents a preliminary exploration of syntactic and semantic aspects of the dialogue in the corpus. We show how simple association metrics can be used to extract words, phrases and more abstract syntactic patterns with targeted meanings or speech-act functions, by making use of the non-linguistic context.

1 Introduction

The use of corpora has proven to be of great value to natural language processing tasks. Parsers for syntactic analysis, for example, have become highly robust and fairly accurate. Advanced semantic processing, however, remains a great challenge. Although applications involving complex use of natural language, such as question answering (Dang et al., 2007), have been shown to profit from deep semantic processing and automated reasoning, a major bottleneck for such techniques, now that several robustness issues have been addressed, appears to be a lack of world knowledge (Giampiccolo et al., 2007). This is not too surprising, since the corpora used are nearly always either text-only or text with some level of, usually task-specific, initially human, annotation. Therefore NLP programs generally have no access at all to non-linguistic context.

A way to get at meaning more naturally is through grounded data and/or grounded interaction, as our own knowledge of natural language meanings is thought to be grounded in action and perception (Roy, 2005). Viewing language as a complex adaptive system which evolves in a community through grounded interaction can yield important new insights (e.g. (Steels, 2003)).

Whereas the techniques for real-world perception in computers are still rather limited, virtual worlds are getting ever more complex and realistic, have many visitors, and do not share the perceptual challenges. This offers great potential for data collection[1]. Examples of virtual word learning-through-interaction projects involving language and/or social behavior are 'Wubble World' (Hewlett et al., 2007) and 'Agent Max' (Kopp et al., 2003).

Our research focuses on learning from data, rather than through interaction, though the latter may be possible in a later stage of the project. We aim at developing algorithms that learn the meanings of words and grammatical constructions in human language in a grounded way. Our data consists of game-logs from the 'Restaurant Game' (Orkin and Roy, 2007), which is an on-line 2-player game in which human players play the roles of customer and waitress in a virtual restaurant. The dataset includes both what they do, and what they say to each other (through chat). It is thus a collection of episodes that take place in a virtual restaurant, enacted by human players, and it has already been shown that useful knowledge about typical activities at restaurants can be extracted from these data. The intuition is that a human student of English starting from scratch (but with some common sense knowl-

[1] von Ahn & Dabbish (2004) were among the first to realize the potential of collecting human knowledge data on-line, in a game setup, collecting a large image-labeling corpus.

edge about how things go in restaurants), could learn quite a bit of English from studying these episodes; possibly enough to play the game. We try to computationally simulate such a learning process. One of the overarching questions underlying this work is what knowledge about language and how it works is needed to extract knowledge about constructions and their meanings from grounded data.

Although the things people say and the things people do tend to be closely related in the restaurant game scenes, the relation is not as straightforward as in some related work, where the data was much more restricted, and the language part contained only descriptions (Gorniak and Roy, 2004) or only directives (Fleischman and Roy, 2005; Gorniak and Roy, 2005). The datasets of those experiments were designed purely for learning word meaning, with each utterance being a nearly direct description of its accompanying action. The Restaurant Game on the other hand was designed for learning natural restaurant behavior, including language, to animate artificially-intelligent characters who can play the game in a convincing, human-like way, and therefore the interaction is much more open-ended. This makes the learning of language from our data a different type of challenge.

In this paper we first introduce The Restaurant Game in section 2, then we explain our main method for extracting words based on their associations with objects in section 3. Next, in section 4, we zoom in on the items on the menu, and extract words and also multi-word units referring to them. This in turn allows us to extract sentence patterns used for ordering food (section 6). Finally we attempt to find words for food items that are not on the menu, by using these patterns, and wrap up with a concluding section.

2 The Restaurant Game

The restaurant theme was inspired on the idea of Schank & Abelson (1977), who argued that the understanding of language requires the representation of common ground for everyday scenarios. Orkin & Roy (2007) showed in The Restaurant Game Project that current computer game technology allows for simulating a restaurant at a high level-of-detail, and exploit the game-play experiences of thousands of players to capture a wider coverage of knowledge than what could be handcrafted by a team of researchers. The goal is automating characters with learned behavior and dialogue. The ongoing Restaurant Game project has provided a rich dataset for linguistic and AI research. In an on-line multi-player game humans are anonymously paired on-line to play the roles of customers and waitresses in a virtual restaurant (http://theRestaurantGame.net). Players can chat with open-ended typed text, move around the 3D environment, and manipulate 47 types of interactive objects through a point-and-click interface. Every object provides the same interaction options: pick up, put down, give, inspect, sit on, eat, and touch. Objects respond to these actions in different ways. For instance, food diminishes bite by bite when eaten, while eating a chair makes a crunch sound, but does not change the shape of the chair. The chef and bartender are hard-coded to produce food items based on keywords in chat text. A game takes about 10-15 minutes to play. Everything players say and do is logged in time-coded text files on our servers. Player interactions vary greatly, and while many players do misbehave, Orkin and Roy (2007) have demonstrated that enough people do engage in common behavior that it is possible for an automatic system to learn statistical models of typical behavior and language that correlate highly with human judgment of typicality.

Figure 1: screen-shot from the Restaurant Game, waitress's perspective

Previous research results include a learned plan-network that combines action and language in a statistical model of common ground that associates relevant utterances with semantic context and a first implementation of a planner that drives AI characters playing the game (Orkin and Roy, 2009).

A total of 10.000 games will be collected, of which over 9000 have been collected already. The average game consists of 85 physical actions and

165 words, contained in 40 lines of dialogue.

Our analyses in this paper are based on a randomly selected set of 1000 games, containing a total of 196,681 words (8796 unique words). This is not a huge amount, but it yields fairly robust results, because we are working with a coherent domain. Of course there will always be utterances that our system cannot make sense of, because sometimes players talk about things that have nothing to do with the game.

The dialogue is grounded in two (partially overlapping) ways. Not only is there a simulated physical environment with objects that can be manipulated in various ways, but also social patterns of reoccurring events provide an anchor for making sense of the dialogue.

3 Associations between objects and words

The game contains a number of different objects, and trying to find words that are used to refer to these is a natural place to start. Let us start out with a simple assumption and see how far it gets us: We expect that objects are most talked about around the times when they are involved in actions. This means we can measure association strength in terms of relative co-occurrence. This is how collocational, or more generally, collostructional strength is commonly measured (Stefanowitsch and Gries, 2003): How often do two things co-occur compared to how often each of them occurs in total? The Chi square (χ^2) value is a good measure of that (Manning and Schütze, 2000).

Game logs were processed as follows. All actions between two lines of dialogue were treated as one action block. All lines of dialogue between two actions were treated as one dialogue block. For each action block it was noted which objects it contains, for each dialogue block which words it contains. Then for each object its association strength with each word was computed based on the occurrence of that word in the blocks of dialogue immediately preceding the blocks containing the object. Preceding dialogue turned out to work better than following dialogue, which might be due to the nature of the corpus with relatively many requests and directives. Only positive associations were taken into account, that is cases where the observed co-occurrence was higher than the co-occurrence that would be expected if words and objects were distributed randomly over the game. In other words, we compare the portion that a word makes up in an object's preceding-block-context to the portion it makes up in the total corpus. The phi value, derived from χ^2 was used as a metric of association strength. We applied basic smoothing (absolute discounting), and required that items occur in at least 4 games in the corpus, to be scored. This reduces noise created by a particular player repeating the same atypical thing a number of times in a game. Table 1 shows all object types with their 5 most strongly correlated words in the preceding dialogue block.

We see that in spite of the simple approach, many objects correlate most strongly with sensible words (we are at this point primarily interested in referring words and phrases). Words for ordered food and drink items are picked up well, as well as those for the menu, the bill, vase and flowers. Some of the kitchen utensils such as the pot and pan are not used often and systematically enough to give good results in this first rough method. When objects are clustered on the basis of their physical interactions, these objects also fail to cluster due to sparse data (Orkin, 2007). The furniture items seem to mostly associate with words for items that are put on them. Looking into the actions in some more detail seems to be needed here, but remains for future work. Relevant context can of course extend beyond the preceding block. We will use a modified notion of context in the next sections.

Since the assumption we made about co-occurrence is so general, we expect it to apply to other domains too: frequently used movable items will most likely pair up with their referring words quite well.

We have observed that in many cases sensible words show up as (most) strongly associated with the objects, but we have no way yet to determine which are the referring ones, which are otherwise related and which are unrelated. Some objects can be referred to by different synonymous words such as 'bill' and 'check'. Others can be referred to by a phrase of more than one word, such as 'spaghetti marinara'. We need to be able to distinguish those cases. The issue is addressed in the following section.

4 Finding words and phrases referring to items on the menu

We will now look in some more detail into the food-items that are ordered (including drinks), listed in table 2. In the present implementation we tell the

object	word 1	phi w1	word 2	phi w2	word 3	phi w3	word 4	phi w4	word 5	phi w5
WATER	**water**	0.24	please	0.02	glass	0.02	thank	0.01	of	0.01
TEA	**tea**	0.34	**te**	0.05	pie	0.02	cup	0.01	**t**	0.01
COFFEE	**coffee**	0.22	coffe	0.03	cup	0.02	tu	0.01	please	0.01
BEER	**beer**	0.26	**beers**	0.03	**berr**	0.02	please	0.02	give	0.02
REDWINE	**red**	0.23	**wine**	0.12	**redwine**	0.02	**wines**	0.02	*glass*	0.01
WHITEWINE	**white**	0.20	**wine**	0.09	**whine**	0.02	red	0.02	degree	0.02
SOUP	**soup**	0.21	**vegetable**	0.05	**jour**	0.04	**de**	0.03	**du**	0.03
SALAD	**salad**	0.17	**cobb**	0.09	cake	0.02	**cob**	0.02	steak	0.02
SPAGHETTI	**spaghetti**	0.18	**spagetti**	0.08	**marinara**	0.04	**pasta**	0.04	steak	0.02
FILET	**steak**	0.25	**filet**	0.14	**mignon**	0.08	lobster	0.03	salad	0.03
SALMON	**salmon**	0.15	**grilled**	0.05	**fish**	0.05	steak	0.01	idiot	0.01
LOBSTER	**lobster**	0.19	steak	0.03	**thermador**	0.03	cake	0.02	salad	0.02
CHEESECAKE	**cheesecake**	0.15	**cake**	0.13	**cheese**	0.08	**cherry**	0.05	**cheescake**	0.05
PIE	**pie**	0.35	**berry**	0.07	cake	0.03	steak	0.02	tea	0.02
TART	**tart**	0.21	**nectarine**	0.08	**tarts**	0.01	coffee	0.01		
MENU	**menu**	0.08	seat	0.03	start	0.02	please	0.02	soup	0.02
BILL	**bill**	0.08	**check**	0.07	pay	0.04	thank	0.03	again	0.02
VASEOFFLOWERS	**flowers**	0.04	these	0.02	**flower**	0.01	**vase**	0.01	roof	0.01
BOWLOFFRUIT	**fruit**	0.05	**fruits**	0.03	**bowl**	0.02	vase	0.01	serious	0.01
BOTTLEOFWATER	**bottle**	0.01	**water**	0.01	cold	0.01	ass	0.01	!	0.01
BOTTLEOFWINE	**bottle**	0.03	**wine**	0.02	brandy	0.02	$50	0.02	dead	0.01
BOTTLEOFBRANDY	**brandy**	0.02	woman	0.02	cake	0.01	whiskey	0.01	road	0.01
BINOFTRASH	**trash**	0.05	**garbage**	0.02	cops	0.02	lmao	0.01	**bin**	0.01
POT	hit	0.02	**pot**	0.02	stuck	0.02	wanna	0.01	yup	0.01
PAN	kitchen	0.01	move	0.01	fish	0.01	an	0.00	off	0.00
MICROWAVE	**microwave**	0.06	kitchen	0.02	break	0.01	staff	0.01	ha	0.01
BLENDER	**blender**	0.01	give	0.01	(	0.01	around	0.01	)	0.01
CUISINART	**blender**	0.03	dropped	0.02	holding	0.02	vase	0.02	out	0.01
CUTTINGBOARD	trash	0.01	pot	0.01	stuck	0.01	wall	0.01	best	0.01
REGISTER	bill	0.07	check	0.06	thank	0.02	no	0.02	pay	0.02
EMPTYTEACUP	**tea**	0.04	refill	0.01	:D	0.01	whenever	0.01	bon	0.01
EMPTYMUG	**coffee**	0.05	check	0.02	**cup**	0.01	thanks	0.01	.	0.01
EMPTYGLASS	beer	0.04	water	0.03	another	0.02	thanks	0.01	thirsty	0.01
EMPTYWINEGLASS	**wine**	0.04	red	0.02	white	0.01	enjoy	0.01	**glass**	0.01
EMPTYBOWL	**soup**	0.03	finished	0.01	entree	0.01	yes	0.01	enjoy	0.01
EMPTYPLATE	enjoy	0.02	else	0.02	dessert	0.02	thank	0.02	anything	0.02
EMPTYWINEBOTTLE	happened	0.02	move	0.01	invisible	0.01	wall	0.01	wonderful	0.01
EMPTYWATERBOTTLE	**bottle**	0.04	they're	0.02	walk	0.01	vodka	0.01	cold	0.01
EMPTYFRUITBOWL	**fruit**	0.02	trash	0.01	serious	0.01	fish	0.01	lol	0.01
EMPTYCUTTINGBOARD	pot	0.01	lol	0.01	**board**	0.01	fish	0.01	best	0.01
EMPTYVASE	**vase**	0.04	flowers	0.03	**flower**	0.02	cost	0.02	they	0.02
EMPTYBRANDYBOTTLE	**brandy**	0.03	asl	0.02	alcoholic	0.02	whiskey	0.01	told	0.01
EMPTYTRASH	**trash**	0.04	woah	0.02	ew	0.02	**garbage**	0.02	flying	0.01
BAR	beer	0.21	water	0.15	wine	0.14	red	0.13	white	0.1
COUNTER	soup	0.06	steak	0.06	salad	0.05	lobster	0.05	tart	0.05
TABLE	please	0.06	water	0.05	wine	0.04	coffee	0.03	soup	0.03
CHAIR	**seat**	0.05	sit	0.04	table	0.04	anywhere	0.03	follow	0.03
STOOL	young	0.02	sup	0.01	wine	0.01	**bar**	0.01	boring	0.01
PODIUM	check	0.08	bill	0.07	else	0.02	no	0.02	the	0.02
MENUBOX	pleae	0.02	hold	0.01	dessert	0.01	second	0.01	minute	0.01
DISHWASHER	microwave	0.02	kitchen	0.01	theres	0.01	look	0.01	that	0.0
STOVE	w	0.01	k	0.01	its	0.01	know	0.00	in	0.00
FRIDGE	cost	0.01	staff	0.01	problems	0.01	vase	0.01	top	0.01
TRASHCOMPACTOR	of	0.01	wine	0.00	go	0.00	the	0.00	water	0.00
BARTENDER	**bartender**	0.03	excuse	0.01	alcoholic	0.01	doin	0.01	mine	0.01
CHEF	favor	0.02	trick	0.02	ha	0.02	ass	0.01	god	0.01

Table 1: all objects types and their 5 most strongly associated words in the preceding dialogue block

system which item types to look at, but automatic object clustering does distinguish food and drink items, too (Orkin, 2007). These items are of interest for a number of reasons. Not only is it highly relevant for the performance of automated characters to be able to recognize which food-items are being talked about when, but they are also interesting because they can be referred to in various ways, and often by expressions consisting of more than one word. Furthermore, there are a number of relevant dialogue acts involving the words for these items, such as ordering. When we can identify the expressions referring to these items, that will also help us identify the environments that these expressions occur in and their function or place in the game.

We will try to extract words and multi-word expressions referring to these objects. In order to avoid all suspicion that we are reproducing the scripted knowledge of the chef and the bartender, we take a slightly different road than before. The point where the customer orders an item is likely to occur earlier than in the dialogue block directly preceding the appearance of the item, or the moment he gets it. So if we want to bypass all interaction with the chef and bartender, it helps to make a rough assumption about where in the game the customer will order, rather than going by our general assumption above. Whereas the above assumption most likely applies to other domains too, this one is a specific assumption based on human knowledge of restaurant scenarios. We cannot make it too specific though, because all games are different.

Every time the waitress puts down a food-item on

a table[2], all customer utterances between this moment and the moment the customer first sat down in the game are considered context for this item. We will refer to this as the order-context for the item. The order-context for an item type is collected by joining the order-contexts of its instances. For the totals we add up the collected order-contexts of all items, rather than taking the totals of the whole corpus. This way we correct for anything that is counted double because it is part of the order-context of more than one item (order-contexts frequently overlap, as in most games more than one item is ordered). The size of this portion of the corpus, without the overlap, is 37,827 words.

4.1 Scoring words and multi-word sequences

Once more we find the most strongly associated words for each item, yielding results similar (but not identical) to table 1. We do the same for two-word and three-word sequences (bigrams and trigrams). For each item we accept the highest scoring word as a good word, assuming that in the minimal case an item can be referred to by exactly one single-word expression. To the extend that our method works, this expression should then be the one that scores highest. Next we accept bigrams that score above a certain threshold if their composing words also score above a threshold (We take phi > 0.02 as a threshold for both). Words that occur in accepted bigrams, but had not been accepted yet, are added to the list of accepted words. Similarly, for trigrams we accept those that score high (same threshold used) and of which the composing bigrams have already been selected in the previous step.[3] The found sequences are presented in table 2.

The approach is somewhat conservative, so we do miss some relevant words, such as 'steak' for FILET (which scored second among the words). We expect that we can catch these later by showing that they occur in the same environments as other food-item expressions. Similarly for the more general words 'fish' and 'pasta' for SALMON and SPAGHETTI respectively, that we saw in table 1. These additionally turn out to have a less strong presence in this part of the data. Presumably they are not used that much in ordering, perhaps because customers, in this situation, tend to repeat what they read on the menu.

item type	unigrams	bigrams	trigrams
WATER	'water'	_	_
TEA	'tea'	_	_
COFFEE	'coffee'	_	_
BEER	'beer'	_	_
REDWINE	'red' 'wine'	'red wine'	_
WHITEWINE	'white' 'wine'	'white wine'	_
SOUP	'soup' 'du' 'jour' 'vegetable'	'soup du' 'du jour' 'vegetable soup'	'soup du jour'
SALAD	'salad' 'cobb'	'cobb salad'	_
SPAGHETTI	'spaghetti' 'marinara'	'spaghetti marinara'	_
FILET	'filet' 'mignon'	'filet mignon'	_
SALMON	'salmon' 'grilled'	'grilled salmon'	_
LOBSTER	'lobster' 'thermador'	'lobster thermador'	_
CHEESECAKE	'cheesecake' 'cherry'	'cherry cheesecake'	_
PIE	'pie' 'berry'	'berry pie'	_
TART	'tart' 'nectarine'	'nectarine tart'	_

Table 2: extracted words, bigrams, and trigrams for the items on the menu

4.2 Filtering referring expressions

We now have identified words and sequences that can be involved in referring to food-items, but we still don't know which of these can be used by themselves for this purpose, and which only as part of a longer sequence. What we do next is to score all words and the selected bigrams and trigrams together in such a way that we only count bigrams where they are not part of one of the selected trigrams and only count the words where they are not part of any of the selected bigrams or trigrams. That is, we treat the bigrams and trigrams selected in the previous step as words, and ignore their internal structure, so we can compare the association scores of these 'words with spaces' to those of other words and in particular with those of their composing words in other configurations. The selected words and bigrams that still score above the threshold now, can apparently refer independently to their associated food-items. This give us the results shown in table 3.

There are two things in this table that are counter-intuitive. Firstly, on the precision side, 'jour' ap-

[2]We could make sure that it is the table the customer actually sits at, but since we only have one customer, the extra administration this would require would probably come with very little gain.

[3]Looking at four-word sequences does not yield additional results if we require that their components have to have been already accepted.

item type	referring expressions
WATER	'water'
TEA	'tea'
COFFEE	'coffee'
BEER	'beer'
REDWINE	'red' 'wine' 'red wine'
WHITEWINE	'white' 'white wine'
SOUP	'soup' 'jour' 'vegetable soup' 'soup du jour'
SALAD	'salad' 'cobb salad'
SPAGHETTI	'spaghetti' 'spaghetti marinara'
FILET	'filet' 'filet mignon'
SALMON	'salmon' 'grilled salmon'
LOBSTER	'lobster' 'lobster thermador'
CHEESECAKE	'cheesecake' 'cherry cheesecake'
PIE	'pie' 'berry pie'
TART	'tart' 'nectarine tart'

Table 3: extracted referring expressions for the items on the menu

Figure 2: menu and specials board from the Restaurant Game

pears to be used outside the expression 'soup du jour' to refer to SOUP, which is quite odd. The most likely cause is that 'du' is relatively often written as 'de', although just not often enough for the whole alternative construction to be picked up (16 times on a total of 79). This issue can be resolved by applying spelling normalization, to recognize that the same word can have different written forms, which will be important to make the final system interact robustly, in any case. As expected in a chat set-up, the spelling is overall rather variable. The opportunities for spelling normalization, however, are promising, since we do not only have linguistic context but also non-linguistic context to make use of. Nevertheless, the theme falls beyond the scope of this paper.

Secondly, on the recall side, 'wine' does not show up as a word that can independently refer to WHITEWINE. Actually, the whole wine situation is a bit particular. Because the word 'wine' occurs prominently in the context of both WHITEWINE and REDWINE it doesn't associate as strongly with either of them as the words 'red' and 'white', which distinguish between the two. In the present implementation our algorithm is not aware of similarities between the two types of objects, which could provide support for the idea that 'wine' is used with the same meaning in both cases. Recognizing and using such similarities remains for future work. It may not seem straightforward either that 'red' and 'white' can refer independently to their objects. What happens is that in the data the word 'wine' can easily occur in a previous utterance of either the customer or the waitress, e.g. waitress: *'would you like some wine?'*, customer:*'yes, red, please.'*. Whether this can be called independent reference is questionable, but at its present level of sophistication, we expect our extraction method to behave this way. Also, because of the medium of chat, players may tend to keep their utterances shorter than they would when talking, using only the distinctive term, when it is clear from the context what they are talking about[4]. Also 'house red/white (wine)' patterns (as appear on the menu in figure 2) do occur in the data but our method is not sensitive enough to pick them up.[5]

In spite of the imperfections mentioned in this step and the previous one (mainly recall issues), we will see in the next section that the expressions we extracted do give us a good handle on extracting patterns of ordering food.

[4]Note that our hard-coded bartender does not respond to the ambiguous order of 'wine' either, as the human designer had the same intuition, that 'red' and 'white' are more reliable.

[5]We are not concerned about not retrieving 'glass of' construction, because we consider it not to be part of the core referring expressions, but a more general construction that applies to all cold drinks.

5 How to order

Now that we have a close to comprehensive collection of expressions referring to food-items, we will use these to find the 'constructions' used for ordering these. For each food-item being put on the table, we record the most recent utterance that contains one of its corresponding referring expressions. We replace this expression by the placeholder '<FoodItem>', so that we can abstract away from the particular expression or its referent, and focus on the rest of the utterance to find patterns for ordering. Table 4 presents the utterance patterns that occurred more than once in a condensed way.[6]

	<FoodItem>	
(and) (a/the/one/another/a glass of/more/some/my)	(and (a/the) <FoodItem>)	(,) (please/plz) (.)
just (a/some)	<FoodItem>	please
yes (,) (a)	<FoodItem>	(please)
(and a)	<FoodItem>	?
glass of/with a/2/um/then/sure	<FoodItem>	
	<FoodItem>	2/too/to start/!
	<FoodItem>	is fine
a	<FoodItem>	would be great
where is my	<FoodItem>	
steak and	<FoodItem>	
	<FoodItem>	
i want	(and a <FoodItem>)	
i'd/i would like (to start with/to have) (a/the/some/a glass of)	<FoodItem>	(please) (.)
i will like	<FoodItem>	
i will start with	<FoodItem>	
i'll take a	<FoodItem>	
	<FoodItem>	
(i think/believe) i'll/i will/ill have (the/a)	(and a glass of <FoodItem>)	(please) (.)
	<FoodItem>	
can/could i have/get (a/the/some/some more/a glass of)	(and (a) <FoodItem>)	(please) (?)
may i have a/the/some	<FoodItem>	(please) (?)
may i please have a glass of	<FoodItem>	?
please may i have the	<FoodItem>	?

Table 4: condensed representation of order-utterances found more than once

There are 492 utterance patterns that occurred more than once, plus another 1195 that occurred only once. Those that occurred twice or more are basically all reasonable ways of ordering food in a restaurant (although some might be the renewal of an order rather than the original one). The vast majority of the patterns that occurred only once were also perfectly acceptable ways of ordering. Many had substantial overlap with the more frequent patterns, some were a bit more original or contained extra comments like *'i'm very hungry'*. We can conclude that there is a lot of variation and that here the extraction method shows a real potential of outperforming hand-coding. As for recall, we can be sure that there are patterns we missed, but also that there will be many possible patterns that do simply not occur in the data. To what extend we will be able to recognize food orders in future games, will largely depend on how successfully we can generalize over the patterns we found.

We envision encoding the extracted linguistic knowledge in the form of a construction grammar (e.g. (Croft, 2001)). The extracted patterns could already be used as very course grained constructions, in which <FoodItem> is a slot to be filled by another construction.[7] At the same time it is clear that there are many recurrent patterns in the data that could be analyzed in more detail. We show initial examples in the subsections 5.1 and 5.2. As for meaning, at utterance level, an important aspect of meaning is the utterance's function as a dialogue act. Rather than describing what happens, most utterances in this game are part of what happens in a similar way as the physical actions are (Searle, 1965). Knowing that something is being ordered, what is being ordered, and how ordering acts fit into the overall scenario will be extremely useful to a planner that drives AI characters.

5.1 Identifying coordination

If we look at sequences associated with ordering, we see that many of them contain more than one <FoodItem> expression. These tend to be separated by the word 'and'. We can support this observation by checking which words are most strongly associated with order phrases that contain 2 or more instances of '<FoodItem>'.The 10 most strongly associated words and their scores are: 'and'(0.19), 'de'(0.05), ','(0.04), 'minon'(0.04), 'i'll'(0.03), '&'(0.03), 'dessert'(0.02), 'the'(0.02), 'with'(0.02), 'n'(0.02). The word 'and' comes out as a clear winner.

Of course 'coordination' is a more general concept than is supported by the data at this point. What is supported is that 'and' is a word that is used to squeeze two <FoodItem> expressions into a single order.

[6] It is worth noting that 97 of the utterances consisted only of '<FoodItem>'. They are included in the first generalized pattern.

[7] Lieven et. al. (2003) argue that young children continue to rely on combining just two or three units well beyond the two-word stage.

5.2 Class-specific constructions

Some of the variation we saw in table 4 is related to there being different types and classes of items that can be distinguished. Table 5 shows this for some relevant classes. This can help us extract more local constructions within the order-phrases and tell the difference between them.

class	trigram	phi
COLDDRINK	'a glass of'	0.06
	'glass of <FoodItem>'	0.06
	'glasses of <FoodItem>'	0.05
HOTDRINK	'cup of <FoodItem>'	0.07
	'a cup of'	0.06
	'<FoodItem> and <FoodItem>'	0.03
STARTER	'<FoodItem> and <FoodItem>'	0.05
	'a <FoodItem> and'	0.04
	'<FoodItem> to start'	0.04
ENTREE	'have the <FoodItem>'	0.07
	'the <FoodItem> and'	0.05
	'and the <FoodItem>'	0.05
DESSERT	'<FoodItem> for dessert'	0.04
	'piece of <FoodItem>'	0.04
	'a piece of'	0.03

Table 5: some interesting classes of item types and their most strongly associated trigrams with phi scores

Here we hand-assigned classes and showed the differences in language, but we could of course start from the other end, and automatically cluster item types on the basis of how they are ordered, thus creating the classes.

6 Finding more words for food-items

It would be great if we could use the knowledge about what ordering looks like, to identify situations where the customer orders something that is not on the menu and figure out how to respond to that (– a challenge because of sparse data).

We extracted the 30 environments of '<FoodItem>', consisting of 2 words to the left and 2 words to the right, that were most strongly associated with ordering, and counted what other words occurred in these environments in the ordering parts of the games. These were the words found, with the number of times they were found in these environments:

'yes'(69), 'menu'(27), 'steak'(16), 'check'(5), 'bill'(5), 'coffe'(4), 'spagetti'(3), 'desert'(3), 'dinner'(3), 'cofee'(2), 'seat'(2), 'fillet'(2), 'sit'(2), 'more'(2), 'dessert'(2), 'you'(2), 'no'(2), 'coke'(2), 'drink'(2), 'bear'(2), 'cute'(1), 'vest'(1), 'help'(1), 'cheese'(1), 'sweet'(1), 'fish'(1), 'ea'(1), 'glass'(1), 'sphagetti'(1), 'burger'(1), 'manager'(1), 'mignon'(1), 'chat'(1), 'cutlery'(1), 'iyes'(1), 'one'(1), 'tab'(1), 'bathroom'(1), 'sieve'(1), 'chesscake'(1), 'selmon'(1), 'med'(1), 'question'(1), 'fast'(1), 'redwine'(1), 'bees'(1), 'bread'(1), 'pudding'(1), 'trash'(1), '?'(1), 'pizza'(1), 'fight'(1), 'cheescake'(1), 'wime'(1), 'wate'(1), 'grilled'(1), 'moment'(1), 'beeer'(1), 'here'(1), '...'(1), 'spegetti'(1), 'pasta'(1), 'spagattie'(1), 'win'(1), 'thank'(1), 'cold'(1), 'main'(1), 'broiler'(1), 'marinara'(1), 'u'(1), 'h'(1), 'refill'(1), 'brandy'(1), 'um'(1), 'whiskey'(1), 'meni'(1), 'acoke'(1), 'cake'(1), 'soda'(1), 'fun'(1), 'offe'(1), 'scotch'(1), 'yours'(1)

These first results look promising and it should not be too hard to filter out misspellings of known words, alternative ways of referring to known food-items, and words that clearly refer to something else known (such as 'menu') (or are simply so frequent that they just have to have some other function). Still, we conclude that the present method on 1000 games is not yet sensitive enough to confidently pick out other food-terms. Improving it remains for future work. This is, on the other hand, a good point to recuperate expressions such as 'steak', which we missed earlier.

7 Discussion/Conclusion

We have picked up all of the menu descriptions for the food-items plus most of the sensible shorter forms. This was good enough to identify patterns of how to order these items.

Our extraction methods have so far been rather human-guided. It would be interesting to see if it is possible to design a more generalized procedure that automatically generates hypotheses about where to look for associations, and what assumptions about the workings of natural language it needs to be equipped with. One basic thing we have used in this case, is the idea that linguistic expressions can be used to refer to things in the non-linguistic context. Another one that is very relevant in The Restaurant Game is that utterances are used as dialogue acts, with very strong parallels to physical actions.

We hope to have given an impression of the richness of this dataset and the possibilities it offers. We argue that finding referring expressions for concrete objects in a simple way is a good starting point in this kind of data to get a handle on more abstract constructions, too.

Acknowledgments

This research was funded by a Rubicon grant from the Netherlands Organisation for Scientific Research (NWO), project nr. 446-09-011.

References

W. Croft. 2001. *Radical Construction Grammar*. Oxford University Press, New York.

H.T. Dang, D. Kelly, and J. Lin. 2007. Overview of the TREC 2007 Question Answering Track. In EM Voorhees and Lori Buckland, editors, *The Sixteenth Text REtrieval Conference Proceedings 2007*, number 500-274 in Special Publication, Gaithersburg, Maryland. NIST.

M. Fleischman and D. Roy. 2005. Why verbs are harder to learn than nouns: Initial insights from a computational model of intention recognition in situated word learning. In *27th Annual Meeting of the Cognitive Science Society, Stresa, Italy*.

D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. 2007. The third PASCAL Recognizing Textual Entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Rochester, New York. Association for Computational Linguistics.

P. Gorniak and D. Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21(1):429–470.

P. Gorniak and D. Roy. 2005. Probabilistic grounding of situated speech using plan recognition and reference resolution. In *Proceedings of the 7th international conference on Multimodal interfaces*, page 143. ACM.

D. Hewlett, S. Hoversten, W. Kerr, P. Cohen, and Y.H. Chang. 2007. Wubble world. In *Proceedings of the 3rd Conference on Artificial Intelligence and Interactive Entertainment*.

S. Kopp, B. Jung, N. Leßmann, and I. Wachsmuth. 2003. Max - A Multimodal Assistant in Virtual Reality Construction. *KI*, 17(4):11.

E. Lieven, H. Behrens, J. Speares, and M. Tomasello. 2003. Early syntactic creativity: A usage-based approach. *Journal of Child Language*, 30(02):333–370.

C.D. Manning and H. Schütze. 2000. *Foundations of statistical natural language processing*. MIT Press.

J. Orkin and D. Roy. 2007. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development*, 3(1):39–60.

J. Orkin and D. Roy. 2009. Automatic learning and generation of social behavior from collective human gameplay. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 385–392. International Foundation for Autonomous Agents and Multiagent Systems.

J. Orkin. 2007. Learning plan networks in conversational video games. Master's thesis, Massachusetts Institute of Technology.

D. Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

R.C. Schank and R.P. Abelson. 1977. *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates Hillsdale, NJ.

J. Searle. 1965. What is a speech act? *Perspectives in the philosophy of language: a concise anthology*, 2000:253–268.

L. Steels. 2003. Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312.

A. Stefanowitsch and ST Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.

L. Von Ahn and L. Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM.