Capturing differences between social activities in spoken language

Jens Allwood
University of Göteborg

1. Introduction

Wittgenstein (1953) introduced the concept of a "language game" and claimed that what we often think of as a language really consists of a collection of more or less diverse "language games". Standard or national languages can therefore only be the result of some form of abstraction performed by linguists (cf. Harris 1980) or other persons, mostly in the service of socio-political interests. It has, therefore, both theoretical and practical interest to investigate to what extent "language games" actually are different and to what extent there exist linguistic phenomena which play an equally important role in all "language games".

Another issue brought up by Wittgenstein's challenge is the question of how we should conceive of "language games". His own remarks are inspiring but require specification to enable empirical investigation of the issues involved. There have been several proposals for how to do this. One proposal is to identify language games with what in the study of literature is called "genres". This is, for example, the approach assumed in Biber (1988). A problem with this approach, if one is mainly interested in spoken language, is that it tends to make analysis of spoken language dependent on what has worked well for written language. Another approach is therefore to take as one's point of departure spoken language. One such approach is the approach suggested in Allwood (1980 and 1995), where variation in spoken language is seen as dependent on what social activity spoken language is serving as an instrument for. The purpose, the roles, the artefacts, environment and domains of the activity are claimed to have a crucial influence on the language (lexicon, grammar and functions) of the activity.

2. Data

To investigate this claim, a spoken language corpus which today comprises about one million words of spoken Swedish has been assembled. The corpus has been collected with the goal of representing many different social activities. The following activity-types (Table 1) are represented in the corpus.

Table 1. The Göteborg spoken language corpus

Activity	Recordings	Words
Shop	13	47337
Occupational therapy	1	8028
Auction	1	12324
Discussion	12	69939
Court	6	33395
Formal meeting	8	156966
Quarrel	1	773
Hotel	8	18931
Informal conversation	13	70738
Interview	45	345169
Classrom interaction	1	3830
Consultation	15	24913
Dinner	3	9872
Trade fair	16	14355
Sermon	2	10311
Radio talk show	2	14086
Role play	3	8138
Conversation in factory	4	24341
Seminar discussion	2	40251
Physical therapy	1	5808
Phone	32	12976
Market	4	12581
TV	2	20222
Task-oriented dialogue	26	15471
Retelling of article	7	5333
Total	226	986088

3. Method

The characterization of the linguistic differences between the different social activities has been done in two ways:

1. manual coding of certain functions of communication
2. development of automatic routines to capture linguistic properties of languages in different social activities

In this paper I will only discuss the automatically derivable properties. What can be derived automatically is, of course, dependent on what is given as input, i.e., on what properties are captured in the transcriptions of our recorded activities. I will therefore briefly present the transcription standard we have used. Consider (1) below:

(1) Transcription according to the MSO4 standard with translation.

§1. Small talk	
$D: säger du de$_0$ ä$_0$ de$_0$ ä$_0$ de$_0$ så$_0$ besvärlit då	$D: oh I see is it it is so troublesome then
$P: ja$_0$ ja$_0$	$P: yes yes
$D: m$_0$ // ha$_0$ / de$_0$ kan ju bli$_0$ så$_0$ se$_1$ du	$D: m // yes / it can be that way you see
$P: < jaha >	$P < yes >
@ <ingressive>	@ <ingressive >
$D: du ta$_1$ den på morronen	$D: you take it in the morning
$P: nej inte på MORRONEN kan ja$_1$ ju tar allti en$_0$ promenad på förmiddan [$_1$ å$_0$]$_1$ då vill ja$_0$ inte ha$_0$ [$_2$ den] $_2$ medicinen å$_0$ sen$_1$ nä$_1$ ja$_0$ kommer hem möjligtvis	$P: no not in the MORNING I always take a walk before lunch [$_1$ and] $_1$ then I don't want [$_2$that] $_2$ medicine and then when I get home possibly
$D: [$_1$ a$_0$] $_1$	$D: [$_1$ yes] $_1$
$D: [$_2$ nä$_0$] $_2$	$D: [$_2$ no] $_2$

As we can see the transcription has the following properties

1. Section boundaries paragraph sign (§). These divide a longer activity up into subactivities. A doctor-patient interview can, for example have the following subactivities: (1) greetings and introduction, (2) reason for visit, (3) investigation, (4) prescribing treatment.
2. Words and space between words.

3. Dollar sign ($) followed by capital letter, followed by colon (:) to indicate a new speaker and a new utterance.
4. Double slash (//) to indicate pauses.
5. Capital letters to indicate contrastive stress.
6. Word indexes to indicate which written language word corresponds to the spoken form given in the transcription (de_0 corresponds to written language det).
7. Overlaps are indicated using square brackets ([ ]) with indices which allow disambiguation if several speakers overlap simultaneously.
8. Comments can be inserted using angular brackets (⟨ ⟩ to mark the scope of the comment and @⟨ ⟩ for inserting the actual comment). These comments are about events which are important for the interaction or about such things as voice quality and gestures.

By using this information, which, in comparison with some transcription systems that provide detailed information about prosody, is relatively close to written language, we have defined a set of automatically derivable properties which include the following:

1. **Volume**: Volume comprises measures of the number of words, pauses, stresses, overlaps, utterances, turns relative to speaker, activity and subactivity.
2. **Ratios:** Various ratios can then be calculated based on the volume measures.

   (2)

mean length per utterance (=MLU)	=	words/utterances
% pauses	=	pauses/words
% stress	=	stress/words
% overlap	=	overlap/words

   Alternatively, pause, stress and overlap can be given per utterance. All of these measures can then be relativized to speaker, activity or subactivity.
3. **Special measures**: One example of a special type of measure is "vocabulary richness" as measured through type/token, Guiraud, Über, Herdan or "theoretical vocabulary", cf. van Hout & Rietveld (1993). Another measure we have constructed is "stereotypicality", which looks at how often words and phrases are repeated in an activity.
4. **Lemma:** We have also implemented a simple stemming algorithm, which enables us to collect regularly inflected forms together with their stems.
5. **Parts of speech**: Parts of speech are assigned using a probability based algorithm and adjusted probabilities originally based on a Swedish written

language corpus tagged for part of speech as input. Words subdivided according to part of speech can then be assigned to speaker, activity or subactivity.

6. **Collocations:** All speakers, activities and subactivities can be characterized in terms of their most frequent collocations.
7. **Sequences of parts of speech:** Utterances of different length can be characterized as to sequence of parts of speech. This allows a first analysis of grammatical differences between speakers, activities and subactivities.
8. **Similarities:** Similarities between activities are captured by looking at the extent to which words and collocations are shared between activities.

4. Examples of the measures applied to six activity types

Let us now consider some examples. They will consist in demonstrating how the measures are applied to six activity types chosen from our corpus (Informal talk at home, Interaction in a computer shop, Doctor-patient consultation, Demonstration at a trade fair, Auction, Sermon in a church).

4.1 Volume and Ratios in the six activities

The percentages are calculated on the basis of the subcorpus created by the six activities (Table 2).

Table 2. Volume of subcorpus

	Informal	Shop	Consultation	Trade Fair	Auction	Sermon
Tokens	70738	41367	18778	14355	12324	9803
Types	6414	4692	2094	2560	907	2265
Utterances	6549	5522	2335	124	154	19
Turns	5594	4725	1991	116	145	17
Tokens %	42.3	24.7	11.2	8.6	7.4	5.9
Utter %	44.5	37.6	15.9	0.8	1.0	0.1
Turns %	44.4	37.5	15.8	0.9	1.2	0.1
MLU	10.8	7.5	8.0	115.8	80.0	515.9

There are also measures of contrastive stress and pausing (Table 3). The percentages are calculated relative to the number of tokens in the corpus.

Table 3. Pauses and stress

	Informal	Shop	Consultation	Trade Fair	Auction	Sermon
Pause	2851	3802	1013	1096	2136	1457
Stress	360	606	116	14	2111	621
Str%Tok	0.5	1.5	0.6	0.1	17.1	6.3
Pau%Tok	4.0	9.2	5.4	7.6	17.3	14.9

Another property directly based on the transcriptions is "overlap of utterances". In Table 4, I give a few measures based on "overlap".

Table 4. Overlap

	Informal	Shop	Consultation	Trade Fair	Auction	Sermon
No of Overl	2461	1276	926	12	18	6
TokOverl	5166	2526	1630	38	28	22
Overl % Utt	37.6	23.1	39.7	9.7	11.7	31.6
Ovetok %Tok	7.3	6.1	8.7	0.3	0.2	0.2

On the basis of the volume and ratio data alone, we can see that the activities fall into two groups. The first group is more interactive and is characterized by more utterances per words, i.e., shorter mean length per utterance (=MLU) and more overlap. The second group is more monological and characterized by fewer utterances and thus by very much longer MLU as well as by less overlap. The sermon and the auction are also characterized by many more pauses and stressed words.

4.2 Special measures

4.2.1 *Vocabulary richness*

Table 5. Vocabulary richness

	Informal	Shop	Consultation	Trade Fair	Auction	Sermon
Type/Token	0.091	0.113	0.112	0.178	0.074	0.231
Über	51.9	51.9	44.1	53.1	34.0	57.7
Herdan	0.785	0.795	0.777	0.820	0.723	0.841
Guiraud	24.1	23.1	15.3	21.4	8.2	22.9
Vocab	399	396	359	448	245	492

We can see that the different measures of vocabulary richness (Table 5) give slightly different results. As has been pointed out in the literature, many of them

are sensitive to text length. The most neutral measure seems to be Vocab (theoretical vocabulary, cf. van Hout & Rietveld 1993). According to this measure, the sermon is richest in vocabulary.

4.2.2 *Stereotypicality*

Table 6. Stereotypicality

Tokens in Collocation	Informal	Shop	Consultation	Trade Fair	Auction	Sermon
1	75.63	74.79	79.95	72.73	88.24	70.27
2	22.93	24.83	28.92	20.11	56.55	16.61
3	3.63	5.16	5.93	3.68	32.16	5.02
4	0.46	0.84	0.99	0.89	17.57	3.06
5	0.07	0.12	0.25	0.36	11.06	2.34
6	0.02	0.03	0.09	0.18	7.68	1.94
7	0.01	0.01	0.02	0.10	5.48	1.71
8	0.01	0.00	0.01	0.06	3.90	1.51
9	0.00	0.00	0.00	0.03	2.70	1.34
10	0.00	0.00	0.00	0.02	1.58	1.19

The measure of stereotypicality (Table 6) shows that the auction is the most stereotypical activity with many repeated collocations. We also see that for collocations with many words all the three activities that involve a greater amount of one-way communication are also more stereotypical. In the case of the sermon, we can see that it is actually the least stereotypical activity on the word and word-pair level, but becomes more stereotypical as the number of words in a collocation increase. This is probably related to the fact that both auctions and sermons involve the use of long standardized phrases.

4.3 Lemmatized Words and Collocations

In order to derive the vocabulary of an activity, we have employed three methods: (1) overall vocabulary frequency, pulling out unique words, (2) only categorematic terms (nouns, verbs, adjectives and adverbs), and (3) lemmatizing categorematic words and collocations. In Table 7, I give a few examples of the third procedure.

Table 7. Lemmatized words and collocations

Informal Words	Shop	Consultation	Trade Fair	Auction	Sermon
åka 'go'	köpa 'buy'	ont 'hurt'	kronor 'crowns'	kronor 'crowns'	gud 'god'
dag 'day'	spelar 'play'	dag 'day'	bilen 'the car'	etthundra 'one hundred'	ande 'spirit'
jobbar 'work'	kronor 'crowns'	tabletter 'pills'	kniven 'the knife'	nummer 'number'	herre 'lord'
år 'year'	kostar 'cost'	barn 'child'	år 'year'	femtiolappen 'the fifty note'	ord 'word'
hör 'hear'	heter 'is called'	titta 'look'	idag 'today'	sjuttiofem 'seventyfive'	heliga 'holy'
Pairs					
tala om 'talk about'	titta på 'look at'	titta på 'look at'	här vise 'this way'	kronor för 'crowns for'	helige ande 'holy spirit'
till exempel 'for example'	som heter 'called'	ont i 'pain in'	års garanti 'year's guarantee'	femtio kronor 'fifty crowns'	den helige 'the holy'
i morgon 'tomorrow'	spel och 'game and'	jag skriver 'I write'	vår bil 'our car'	såld för 'sold for'	vi ber 'we pray'
i dag 'today'	köpa en 'buy a'	på morgonen 'in the morning'	kronor för 'crowns for'	femtiolappen för 'fifty note for'	jesus kristus 'Jesus Christ'
fråga om 'ask about'	heter det 'is called'	i benet 'in the leg'	kniv som 'knife which'	åttiosju nummer 'eighty seven items'	ber för 'pray for'
Triples					
vad heter det 'what's it called'	spel och sådant 'games and such'	det gör ont 'it hurts'	det här vise 'this way'	kronor för den 'crowns for it'	den helige ande 'the holy spirit'
och hälsade på 'and greeted'	titta på den 'look at it'	är du rädd 'are you afraid'	under dagens hushåll 'during the fair'	femtio kronor för 'fifty crowns for'	vi ber för 'we pray for'
över huvud taget 'at all'	vad heter det 'what's it called'	vi titta på 'we look at'	på detta viset 'in this way'	viktor åttiosju nummer 'viktor eightyseven number'	låt oss bedja 'let us pray'

kommer du ihåg 'do you remember'	lämnade in den 'delivered it'	titta på det 'look at it'	på det viset 'in this way'	femtiolappen för den 'fifty note for it'	hör vår bön 'hear our prayer'
är frågan om 'a question of'	ha en påse 'have a bag'	och titta på 'and look at'	vaxar din bil 'wax your car'	femtiolappen på den 'fifty note on it'	fadern och sonen 'the father and the son'

The translations fairly clearly indicate that the distinct domains of the six activities are reflected in the words and collocations that are selected. The reason for this is that many activities coincide with particular conceptual domains. We could talk about religion in a shop, but mostly we don't, so that even when there is no necessary link between a particular activity and a particular conceptual domain, there might be such a link in practice and this will be empirically observable.

4.4 Parts of Speech

In Table 8, data on parts of speech in the six activities are presented. The table also gives data on the total number of tokens in the activities and the relative share of the parts of speech in all activities. Note that we have introduced "feedback words" (fb), "own communication managent words", e.g. hesitation sounds as a parts of speech. There are also the two categories of "phrases which function as single units" (phras) and "delimiters" (del) which are the signs we use to mark pauses, overlaps and comments with. See the description of the transcription conventions above.

Table 8. Parts of speech

	Informal	Shop	Consul-tation	Trade Fair	Auction	Sermon	Tot%	Tot
pron	22.96	22.16	22.32	15.72	16.33	13.76	21.04	35214
verb	20.71	19.89	20.82	16.90	11.95	14.53	19.19	32110
adv	17.77	15.65	17.93	24.50	22.36	21.78	18.41	30818
noun	12.01	12.53	11.15	18.06	16.31	23.73	13.56	22701
conj	8.09	6.12	7.87	6.76	3.13	8.19	7.10	11888
prep	6.56	5.49	5.63	3.66	6.43	2.75	5.71	9554
fb	6.45	7.56	7.98	1.62	0.88	0.20	5.71	9553
adj	3.38	3.00	2.98	8.62	2.67	12.14	4.15	6946
num	0.80	1.67	0.63	2.24	16.96	1.42	2.34	3923
del	0.62	4.17	1.76	1.65	2.45	0.18	1.82	3054
ocm	0.43	1.15	0.83	0.10	0.49	1.12	0.67	1118
int	0.23	0.61	0.11	0.16	0.02	0.19	0.29	479
phras	0.00	0.01	0.00	0.00	0.00	0.00	0.00	7
Tot%	42.27	24.72	11.22	8.58	7.36	5.86	100.0	
Tot	70738	41367	18778	14355	12324	9803		167365

As we can see, pronouns are totally the most frequent part of speech, followed by verbs and adverbs. This is a fairly stable result for spoken language. This is also, perhaps not surprisingly, the order we find in the three first activities which are the most interactive. In the three last activities, the order is different. In the Trade fair, adverbs and nouns are the two most frequent. This can probably be explained by the nature of the activity where a person is demonstrating various objects to an audience. In the Auction, adverbs and numerals are the most frequent and pronouns and nouns are third and fourth most frequent. Again, this can probably be explained by the nature of the activity, which involves displaying and bidding for displayed objects. Finally, the sermon has nouns and adverbs as the most frequent parts of speech. It also has an unusually high share of adjectives — higher than any other activity. It is likely that these features can be related to the role that biblical quotations and descriptions play in this activity.

Let us now turn our attention to more particular data about each part of speech. The most common words in four parts of speech are presented for each activity in Table 9. Of the three parts of speech, nouns show the most interesting differences between the activities. (English translations are given only the first time that a word occurs in the tables).

The table shows that syncategorematic function words, e.g., pronouns and

Table 9. Pronouns, verbs, nouns and feedback

Informal	Freq	Shop	Freq	Consultation	Freq	Trade Fair	Freq	Auction	Freq	Sermon	Freq
Pronouns											
det 'it'	4221	det	2519	det	1309	det	650	vi	768	det	213
jag 'I'	2254	jag	1017	du	657	vi	291	den	392	vi	149
du 'you'	1271	du	765	jag	633	ni 'you'	168	det	219	oss 'us'	86
de 'they'	1013	den	742	man	168	den	163	en	129	han 'he'	63
man 'one'	765	vi	473	den	152	en	109	vem 'who'	108	dig 'you'	52
den 'it'	701	en	473	en	137	man	101	ingen 'no one'	84	din 'your'	51
en 'a'	676	de	443	de	113	du	83	jag	82	den	51
vi 'we'	614	som	344	som	105	jag	75	ett 'a'	43	som	49
han 'he'	561	man	284	vi	103	som	68	alla	27	dem 'them'	49
som 'that'	541	vad 'what'	279	vad	78	alla 'all'	52	någon 'some-one'	25	honom 'him'	48
Verbs											
är 'is'	1756	är	1290	är	511	är	366	har	470	är	148
har 'has'	937	har	629	har	338	har	204	är	147	har	83
var 'was'	805	ska	335	ska	170	kan	142	hade	142	älskar 'loves'	33
kan 'can'	521	kan	316	var	157	ska	119	börja 'begin'	112	kan	29
ska 'will'	454	var	273	kan	154	tar	57	kommer	109	kommer	28
vet 'knows'	354	ha	227	ha	118	ta	50	börjar 'begins'	71	hade	27
får 'gets'	331	får	167	gör 'does'	95	gör	46	går 'walks'	66	ska	26
hade 'had'	325	vet	164	får	86	kommer 'comes'	45	ger 'gives'	59	vill 'wants'	22
skulle 'should'	315	tror 'thinks'	125	ta 'take'	78	blir 'becomes'	38	ska	57	ber 'prays'	22
ha 'have'	248	köpa 'buy'	112	tar 'takes'	66	vet 'knows'	31	får	52	hör 'hears'	21

Nouns

fall 'case'	123	tack 'thanks'	175	dag 'day'	49	man 'one'	128	kronor 'crowns'	336	gud 'god'	103
dag 'day'	119	kronor 'crowns'	84	tabletter 'tablets'	48	kronor 'crowns'	51	nummer 'number'	163	ande 'spirit'	67
år 'year'	101	fall 'case'	60	benet 'the leg'	39	bil 'car'	49	femtiolappen 'the fifty note'	162	herre 'lord'	55
gång 'time'	90	spel 'game'	43	vecka 'week'	36	kniven 'the knife'	43	hundralapp 'hundred note'	78	ord 'word'	48
vecka 'week'	52	datorn 'the computer'	40	gång 'time'	34	år 'year'	24	lådan 'the box'	58	fadern 'the father'	38
jobbet 'the job'	46	mats	35	blodtryck 'blood pressure'	31	gång 'time'	23	viktor (name)	44	kyrka 'church'	35
klockan 'the time'	45	vecka 'week'	32	år 'year'	28	viset 'the way'	22	sigvard (name)	31	människor 'people'	33
tid 'time'	43	spänn 'buck'	31	doktorn 'the doctor'	26	munstycke 'mouth-piece'	21	oljan 'the oil'	27	man 'man'	32
morgon 'morning'	42	bilden 'the picture'	31	medicinen 'the medicine'	25	dag 'day'	21	lars (name)	24	kärlek 'love'	30
mats (name)	42	priset 'the price'	29	ögat 'the eye'	22	burken 'the jar'	20	vatten 'the water'	23	jesus (name)	30

Informal	Freq	Shop	Freq	Consultation	Freq	Trade Fair	Freq	Auction	Freq	Sermon	Freq
Feedback											
ja 'yes'	2252	ja	1453	ja	867	ja	110	ja	90	ja	13
nä 'no'	700	nä	475	m	176	va	84	jaha	7	va	1
m	468	m	409	nä	142	jo	9	nä	6	nä	1
va 'what'	294	va	302	nej	98	nej	8	niao	2	m	1
hm	222	okey	107	jaha	45	nä	7	va	1	jo	1
jo 'yes'	143	jaha 'yes'	92	va	42	okey	3	jo	1	javisst 'yes sure'	1
jaja 'yesyes'	111	nej	56	jo	29	ne 'no'	2	jaså	1	a 'yes'	1
nej 'no'	91	jo	52	hm	25	m	2	jae 'yes'	1		
jaså 'I see'	73	jaså	41	jaså	22	a	2				
jaja 'yesyes'	57	jaja	17	aha	11	nähä	1				

feedback words, are shared to a much greater extent than categorematic words, e.g., most verbs and nouns. Only in the case of nouns is there a big difference, in accordance with the activity domains, between what words are frequent in the six activities. The verbs do not show an equal differentiation, since most of the highly frequent verbs are auxiliaries, i.e., syncategorematic.

5. Sequence of parts of speech in specific utterance types

Another measure of distinctiveness can be obtained by studying what sequences of parts of speech are typical of each activity. Space allows me to exhibit only the sequences of utterances found in the shop activity (Table 10). As we can see, these utterances mostly are what can be expected in a shop context.

Table 10. Collocations
Shop

Part of speech	Frequency	Most common example	Freqency of example
1-word utterances			
feedback (=fb)	675	m_0 'm'	194
noun	151	tack 'thank you'	45
int	95	hej 'hi'	70
adv	39	$så_0$ 'so'	15
verb	35	ha_0 'yes'	23
ocm	34	m_1 'm'	26
num	20	hundra 'hundred'	3
pron	18	vadå 'what'	3
adj	9	mellanvänd 'turned between'	2
conj	3	$å_0$ 'and'	1
2-word utterances			
fb fb	87	ja_0 ja_0 'yes yes'	20
noun noun	33	jonn silver 'john silver'	2
fb int	27	ja_0 tack 'yes thanks'	8
fb adv	21	ja_0 visst 'yes certainly'	6
verb pron	16	sa_0 du 'did you say'	2
3-word utterances			
pron verb pron	37	va_1 sa_0 du 'what did you say'	10
fb adv pron	21	a_0 just de_0 'yes precisely'	11
fb fb fb	18	$nä_0$ $nä_0$ $nä_0$ 'no no no'	3
pron verb adv	15	ja1 vet inte 'I don't know'	1
int adv adv	15	tack $så_0$ mycke 'thanks a lot'	12

4-word utterances			
verb pron adv adv	19	va_3 de_0 bra $så_0$ 'was it good like that'	9
fb pron verb pron	18	$nä_0$ ja_1 har de_0 'no I have that'	1
fb pron verb adv	14	näe han sa_0 ju 'no he said you know'	1
pron verb pron noun	10	vi_0 bytte en_0 grej 'we changed a thing'	1
pron verb pron adv	10	du tror de me 'you think so too'	1
5-word utterances			
fb pron verb pron adv	9	okej ja_1 tar de_0 då 'ok I will take that then'	1
pron verb pron prep pron	5	va_1 tar ni för dom_2 'what do you charge for them'	1
adv verb pron verb pron	4	$så_0$ kan ja_1 ta_0 dom_2 'so can I take them'	1
verb pron verb pron noun	3	$ä_0$ de_0 e_0 nån häst 'is that a horse'	1
verb pron adv adv adv	3	$ä_0$ re_0 $hä_1$ bra då 'is this good then'	1
6-word utterances			
verb pron verb pron noun adv	3	ska du $få_0$ en_0 påse också 'would you like a bag too'	1
fb pron verb adv verb pron	3	m_0 man kan nästan tro_0 de_0 'you can almost believe that'	1
fb pron verb adv prep noun	3	okej ja_1 kikar in i morron 'ok I will look in tomorrow'	1
fb pron fb pron pron verb	3	nähe de_0 va_0 de_0 ja_1 trodde 'no that's what I thought'	1
7-word utterances		ja_0 de_0 ska den ju säkert göra	
fb pron verb pron adv adv verb	2	'yes it should certainly do that'	1
fb pron verb adv verb pron adv	2	$nä_0$ ja_1 vet inte va_3 de0 va_3 'no I don't know what it is'	1

6. Similarities between activites

Let us now turn to the question of the extent to which words and collocations are shared between activities (Table 11). Only words and pairs are translated. In Swedish spoken language, it seems as if constructions involving *de* (it) as a topic marker with a copula *e* (is) or a conjunction *å* (and) are among the linguistic *sine qua non* of most social activities.

Table 11. Words and collocations

Number of activities	Relative frequency rank	Totfreq	Words
6	5.9812	8187	de_0 'it'
6	5.9725	3839	$å_0$ 'and'
6	5.9684	3913	$så_0$ 'so'
6	5.9600	2489	på 'on'
6	5.9554	2567	den 'it'
Pairs			
6	5.8534	1431	de_0 e_0 'it is'
6	5.4405	552	e_0 de_0 'is it'
6	5.3953	157	på de_0 'on it'
6	5.3474	454	de_0 va_3 'it was'
6	5.2874	447	$å_0$ $så_0$ 'and so'
Triples			
6	4.2638	49	$å_0$ de_0 e_0 'and it is'
6	4.0500	115	ja_0 de_0 e_0 'yes it is'
6	3.2034	22	nu ska vi_0 'now will we'
5	4.1206	203	de_0 e_0 ju 'it is you know'
5	4.0764	143	i alla fall 'in any case'
Quadruples			
4	2.1437	18	de_0 e_0 de_0 som 'it is what which'
4	2.0090	17	de_0 e_0 ju en_0 'it is a'
4	1.9291	15	då ska vi_0 se_0 'then let us see'
4	1.8027	14	ja_0 de_0 e_0 de_0 'yes it is it'
4	1.8027	14	de_0 e_0 inte $så_0$ 'it is not so'

Quintuples			
3	1.0595	7	ja_1 vet inte om_0 de_0 'I doní know if it'
3	1.0343	6	de_0 e_0 de_0 som e_0 'it is what which is'
3	0.9336	5	de_0 e_0 ju $så_0$ att 'it is you know so that'
3	0.8819	3	ja_1 vill att du ska 'I want you to'
3	0.8604	4	i $å_0$ me_0 att du 'since you'

We have also developed a measure of the relative amount of words and collocations occurring in one or more of the activities in a corpus. This is presented in Table 12.

Table 12. Percentage of words and collocations occurring in a given number of activities

Length of coll.	Tok	Type	non-uniq	% in 1	% in 2	% in 3	% in 4	% in 5	% in 6
1	164310	12970	3089	76.18	12.34	5.32	2.85	2.16	1.15
2	148879	65939	8489	87.13	7.99	2.91	1.39	0.47	0.11
3	137146	110385	4697	95.74	3.33	0.71	0.18	0.03	0.00

The table shows that 76% of all words are confined to only one activity, 12 % occur in two activities, 5% in three activities etc. Only a very small part of the vocabulary, 1%, is shared between all six activities. Turning to collocations, we see that they are, to an even greater extent, connected with a single activity. This is so because most of the collocations only have a frequency of one.

7. Discussion

This paper presents a number of ways in which linguistic properties of distinct social activities can be derived from transcriptions made according to a format with a relatively small number of special requirements.

The properties extracted show that the activities differ from each other in volume and various ratios. MLU is an example of a ratio that directly measures mean length of utterance but can also be seen as a measure of interactivity. There are also differences as measured by "vocabulary richness" and "stereotypicality". Furthermore, there are differences in vocabulary and in the parts of speech which are typically employed in an activity. Finally, we have been able to demonstrate differences in the collocations and in the grammatical structures

as indicated by the sequences of parts of speech which are typical of utterances with a given length. Turning to similarities, we have been able to show that certain words and collocations seem to be employed in all the activities that were studied.

Returning to the original question posed in this paper, it seems that language to some extent can be seen as a collection of "language games", which are distinct from each other in vocabulary, structure and function. But there also seem to be a very small number of certain highly frequent words and structures used in all activities. By and large, these words are all syncategorematic or function words, i.e., words helping to structure what we say.

One question that this raises is whether the differences between the activities can be compared to differences between "national standard languages". Can whole national standard language communities be characterized by some of the measures employed above? Could there be, for example, typical patterns of MLU, or typical uses of pauses and overlaps, preferred levels of vocabulary richness and word frequencies which differentiate national standard languages? For some of these measures, we have reason to believe that this is so. Concerning overlaps, for example, Fant (1995) have demonstrated that Swedish has a much lower rate than Spanish in the activity of business negotiation — a difference between Swedish and Spanish that might be repeated in other activities. Concerning word frequencies, it is also rather well-known that there are interesting differences. It remains an open question, however, if the range of linguistic variation between the activities of a single national language community is as great as what can be found between different national languages.

References

Allwood, Jens

1980 "On Power in Communication". In *ALVAR: a Festschrift to Alvar Ellegård* [*SPELL* I], J. Allwood and M. Ljung (eds), 1–20. University of Stockholm, Department of English.

1995 "An Activity Based Approach to Pragmatics". *Gothenburg Papers in Theoretical Linguistics* 76: 1–38.

Biber, Douglas

1988 *Variation Across Speech and Writing.* Cambridge: Cambridge University Press.

Fant, Lars

1995 "Negotiation discourse and interaction in a cross-cultural perspective: The case of Sweden and Spain". In *The Discourse of International Negotiations*, K. Ehlich and J. Wagner (eds), 177–201. Berlin-New York: Mouton de Gruyter.

Harris, Roy
1980 *The Language Makers.* London: Duckworth.
Van Hout, Roeland and Rietveld, Toni
1993 *Statistical Techniques for the Study of Langauge and Language Behaviour.* Berlin & New York: Mouton de Gruyter.
Wittgenstein, Ludwig von
1953 *Philosophical Investigations.* Oxford: Basil Blackwell.