Conversational types: a topological perspective

Kwong-Cheong Wong & Jonathan Ginzburg
Laboratoire de Linguistique Formelle (UMR 7110)
Université Paris-Diderot (Paris 7)
LabEx-EFL, Sorbonne Paris Cité, France
wongkwongcheong@gmail.com

Abstract

The notion of conversational genre/type is a crucial one for various tasks in dialogue. These include the planning of the subject matter of initiating utterances, the form/content of domain-specific moves, and the resolution of non-sentential utterances. In this paper, we discuss experiments whose aim is to come up with metrics over the class of conversational types. We compare two main methods: using n-grams (n=1,2) and using the distribution of non-sentential utterances. We show that both methods yield promising results, though the method involving non-sentential utterance distributions is ultimately more effective. We consider the implications that this has for modelling conversational types.

1 Introduction

The notion of a *language game* (Wittgenstein, 1953) or a *speech genre* (Bakhtin, 1986) is one of the most fundamental in research on dialogue. We will use the term *conversational type*, henceforth. There has been intermittent work on this notion in the pragmatics literature: Hymes (1972) suggests that a conversational type can be characterized by eight parameters SPEAKING – Scene, Participants, Ends, Act sequence, Key, Instrumentalities, Norms and Genre; Levinson (1979) takes such a notion to 'refer to a fuzzy category whose focal members are goal-defined, socially constituted, bounded events', and proposes three dimensions according to which activity types vary: *scriptedness* (the degree to which the activity is routinized), *verbalness* (the degree to which talk is an internal part of the activity) and *formality* (the degree to which the activity is formal or informal). For instance, teaching is much more verbal than a football game, and a jural interrogation is both much more formal and scripted than a dinner party. Allwood (1995) proposes that such a notion can be further characterized by four parameters: *purpose of the activity*, *roles performed by participants*, *instruments used*, and *other physical environment*. Schank and Abelson (1977) argue that most of human understanding is script-based. A *script* is a way of representing what they call "specific knowledge," that is, detailed knowledge about a situation or event that "we have been through many times." (p.37) A script consists of various *slots* to be filled by different elements according to that particular script. The general idea underlying this notion, then, relates to what an agent needs to learn in order to participate successfully in a given conversational type. From a concrete point of view of dialogue modelling, the role played by conversational types as the basis for explaining domain specificity includes *at least* three aspects we exemplify here with constructed examples:

1. Special forms usable at particular points and their non-sentential meanings, e.g., with respect to opening/closing interaction:

   (1) a. A: Hi. B: Hi. (A and B go their separate ways).
   b. The court is now in session. . . . This session is now closed.
   c. A: Welcome to today's auction. . . . That brings us to the end of today's auction.

   (2) a. Initially: Umpire: player X to serve, love all.
   b. During game: Umpire: X-Y (=Server has X points, receiver has Y points)
   c. At end of game: Umpire: game Z (=Player Z has won the game)

2. Non-locally determined relevance:

   (3) a. (First utterance in a bakery:) A: Two croissants.
   b. Initial stage of informal chat between A and B: A: How are you? How is the family?

3. Conversational completeness: when can a conversation be considered to have achieved its goals which allows the participants to terminate it.

Building on earlier AI work on planning (e.g., (Cohen and Perrault, 1979; Litman and Allen, 1987)), Larsson (2002) models plans as sequences of questions; domains are distinguished by the sets of questions whose resolution is required. This provides the basis for the family of systems following Godis (Larsson and Berman, 2016). Within the framework of KoS, Ginzburg (2012) proposes to model a conversational type in terms of a type that characterizes the information state of a participant that has *completed* a conversation of that kind. On this view, a conversational type directly specifies information about the participants (including potentially relationships that hold between them), the subject matter (via the field QNUD (questions no longer under discussion)), and certain moves:

(4)
$$\begin{bmatrix} \text{participants} : \begin{bmatrix} \text{person1} : \text{Ind} \\ \text{cperson1} : \text{cp1(person1)} \\ \text{person2} : \text{Ind} \\ \text{cperson2} : \text{cp2(person2)} \end{bmatrix} \\ \text{qnud} : \text{poset(question)} \\ \text{moves} : \text{list(utterance-type)} \end{bmatrix}$$

There is, thus, conceptual and formal work on conversational types, some of which has been implemented. However, due to its symbolic nature, basic topological notions relating the closeness/similarity between types have not hitherto be considered. Nor have there been attempts at characterizing the global structure of the space of conversational types. This is presumably an open class, but by analogy with the lexicon, plausibly possesses internal structure—, say, a subclass of types that allow for relatively free interaction or ones where some participants are essentially silent etc.

In this paper, we describe experiments whose aim is to develop basic topological notions on a given ensemble of conversational types. Our aim is to develop computational techniques that enable us to diagnose automatically for a new conversational type its location in relation to other conversational types. We do this by defining a metric between types on the basis of several distinct probability distributions:

(5) A *metric* on a set X is a function (called the distance function) $d : X \times X \mapsto \mathbb{R}^+$ (where $\mathbb{R}^+$is the set of non-negative real numbers) that satisfies (i) symmetry: $d(a, b) = d(b, a)$, (ii) identity: $d(a, b) = 0$ if and only if $a = b$, (iii) non-negativity: $d(a, b) \geq 0$, and (iv) the triangle inequality: $d(a, c) \leq d(a, b) + d(b, c)$.

We use the Jensen-Shannon divergence (JSD), which is a metric created from the Kullback-Leibler (KL) divergence measure. In (6) P and Q are two given probability distributions:[1]

(6) a. KL divergence $D(P||Q) =_{def} \Sigma_i p(i)\ log\ p(i)/q(i)$

b. $JSD(P|Q) = .5D(P||M) + .5D(Q||M)$ with $M = .5(P + Q)$

As a set of conversational types we take the BNC (British National Corpus) taxonomy (Burnard, 2000). We consider two main approaches: in section 2, we use *n*-grams $(n = 1, 2)$, the intuition being that this involves clustering on the basis of 'subject matter'; in section 3, we use the distribution of non-sentential utterances, the intuition being that this involves clustering on the basis of 'interactional structure', as we explain below. In section 4, we offer a comparative evaluation of the two approaches, the impact of which is discussed in section 5. Finally, in section 6 we draw some conclusions and suggest future work.

2 Metrics using unigrams and bigrams

2.1 Experimental details for unigrams

We obtained the 23 unigram frequency files, one for each of the 23 (classified) BNC spoken genres from the BNC*web* (CQP-Edition)[2], restricting the POS-tags to any verb and any noun. (For the names and descriptions of these 23 BNC spoken genres, see Table 1).

[1] In fact, JSD as defined here is the square of a metric Fuglede and Topsoe (2004).

[2] http://bncweb.lancs.ac.uk/

Genre	Description	Genre	Description
1 Broadcast_Discussion (**Discn**)	TV or radio discussions	13 Lecture_Natural_Science (**Nat_sc**)	lectures on the natural sciences
2 Broadcast_Discussion (**Doc**)	TV documents	14 Lecture_Politics_Law_Education (**P_law**)	lectures on politics, law or education
3 Broadcast_News (**News**)	TV or radio news broadcasts	15 Lecture_Social_Science (**Soc_sc**)	lectures on the social sciences
4 Classroom (**Class**)	non-tertiary classroom discourse	16 Meeting (**Meet**)	business or committee meetings
5 Consultation (**Cons**)	mainly medical consultations	17 Parliament (**Prlmnt**)	parliamentary speeches
6 Conversation (**Conv**)	face-to-face spontaneous conversations	18 Public_Debate (**P_deb**)	public debates and discussions
7 Courtroom (**Court**)	legal presentations or debates	19 Semon (**Sermn**)	religious sermons
8 Demonstration (**Demo**)	'live' demonstrations	20 Speech_Scripted (**Sp_s**)	planned speeches
9 Interview (**Intv**)	job interviews and other types	21 Speech_unscripted (**Sp_us**)	uplanned speeches
10 Interview_Oral_History (**Hist**)	oral history interviews	22 Sportslive (**Sport**)	'live' sports commentaries and discussions
11 Lecture_Commerce (**Comm**)	lectures on commerce	23 Tutorial (**Tut**)	university-level tutorials
12 Lecture_Humanities_Arts (**H_arts**)	lectures on humanities and arts subjects		

Table 1: BNC spoken genres (Hoffmann et al., 2008) p. 276

Following common practice in text categorization, stop words (functions words and other uninformative words) were then filtered out from these files. The set of stop words we used was the one provided by the free statistical software R (R-Core-Team, 2013) (174 in total) as shown in Table 13 in the Appendix, plus the following 20: *'ve*, *'s*, *'re*, *'m*, *'ll*, *'d*, *d'*, *sha*, *wo*, *can*, *ca*, *will*, *must*, *may*, *might*, *shall*, *shalt*, *used*, *need*, *dare*. Note that there is no universal set of stop words and researchers have used different sets of stop words (Manning and Schütze, 1999), usually tailor-made to their specific tasks. The size of the set of stop words we used (194) is minimal as compared to those of the others (e.g., 527 in Weka (Witten et al., 2016)). We believe that a minimal set of stop words is likely to be more appropriate to our present study as there are 23 different spoken genres and stop words in some genres may not be stop words in the other genres.[3] From each of the 23 filtered unigram files, we selected its top 100 most frequent unigrams, and then obtained the union set of these 2,300 unigrams by amalgamating these and deleting duplicates. The resulting union set contained just 821 unigrams in total, whose 50 most frequent members are shown in Table 2.

From the perspective of vector space models (Clark, 2015), these 821 selected unigrams result in an 821-dimensional vector space with each selected unigram representing one dimension. Each of the 23 genres is represented by a point (or vector) in this higher dimensional vector space. The position of each genre-point is determined by the probability distribution of the 821 selected unigrams in the genre in the following way: the magnitude along the dimension represented by the selected unigram is given by the value of the probability of occurrence of that selected unigram among the 821 selected bigrams in the genre. The latter is the ratio of the normalized frequency of that unigram in the genre to the total normalized frequency of the 821 selected unigrams in the genre.The distance between each and every pair of genre-points is then measured using Jensen-Shannon Divergence (JSD), as defined in section 1. Figure 1 displays this data using a force-directed graph (FDG) (Bannister et al., 2012). The distance matrix for this metric sorted by closest neighbour is displayed in full in Table 10 in the Appendix.

Rank	Unigram	Rank	Unigram	Rank	Unigram	Rank	Unigram	Rank	Unigram
1	know	11	mean	21	take	31	done	41	day
2	think	12	way	22	thing	32	fact	42	number
3	got	13	said	23	bit	33	mr	43	saying
4	get	14	want	24	point	34	year	44	god
5	people	15	come	25	work	35	use	45	end
6	say	16	sort	26	course	36	says	46	thought
7	see	17	put	27	lot	37	gonna	47	went
8	go	18	things	28	give	38	find	48	case
9	going	19	look	29	years	39	made	49	tell
10	time	20	make	30	like	40	government	50	week

Table 2: 50 most frequent unigrams in the union set

2.2 Experimental details for bigrams

There are different ways to extract bigrams in the literature (e.g., (Tan et al., 2002)). We used the software AntConc (Anthony, 2017) to extract bigrams from the text files of the 23 BNC spoken genres.

[3] In fact, stop words are not always used in experiments in other fields, such as register variation in applied linguistics (see, e.g., Biber and Egbert (2016)). In order to investigate the effects of using stop words on our results, we repeated our experiments without using stop words. We obtained no significantly different results. Due to space constraints, we report herein only the results of experiments that used stop words.

Following common practice, we filtered cases where either component of the bigram is a stop word from the extracted bigrams, using the same set of stop words we used for unigrams above. As with the unigrams, we selected from each of the 23 filtered bigram files its 100 most frequent bigrams, and then obtained the union set of these 2,300 bigrams by amalgamation and deletion of duplicates. The resulting union set contained 1410 bigrams in total, whose 50 most frequent members are shown in Table 3. The same procedure was then followed to generate the JSD metric of the bigram distributions of the 23 BNC spoken genres. Figure 2 displays this data using an FDG. The distance matrix for this metric sorted by closest neighbour is displayed in full in Table 11 in the Appendix.

Rank	Bigram	Rank	Bigram	Rank	Bigram	Rank	Bigram	Rank	Bigram
1	er er	11	oh yes	21	twenty five	31	yeah erm	41	first time
2	yeah yeah	12	right now	22	two hundred	32	three hundred	42	one thing
3	yes yes	13	come back	23	united states	33	oh right	43	two thousand
4	little bit	14	five percent	24	yeah well	34	right erm	44	er well
5	erm er	15	last year	25	greater york	35	erm well	45	one point
6	something like	16	go back	26	new settlement	36	right yeah	46	thought
7	right okay	17	years ago	27	next week	37	thousand pounds	47	jesus christ
8	nineteen eighty	18	things like	28	o clock	38	say well	48	one hundred
9	county council	19	mm mm	29	oh yeah	39	labour party	49	long time
10	nineteen ninety	20	make sure	30	last week	40	nineteen forty	50	er erm

Table 3: 50 most frequent bigrams in the union set

3 A Metric based on NSU distributions

Corpus studies of non-sentential utterances (NSUs), a characterizing feature of dialogue—fragments which express a complete meaning—show that 'sentential' fragments can be reliably classified using a small, semantically-based taxonomy (Fernández and Ginzburg, 2002; Schlangen, 2003). In the taxonomy of Fernández and Ginzburg (2002), for instance, which attains high coverage of a large random sample of the BNC (98.9%), there are 15 classes of NSUs, covering various kinds of acknowledgments (plain acknowledgement, repeated acknowledgement), queries (clarification ellipsis, sluice, check question), answers (short answer, plain affirmative answer, repeated affirmative answer, propositional modifier, plain rejection, helpful rejection), and extensions (factual modifier, bare modifier phrase, conjunction + fragment, filler); see Table 4 for examples. The taxonomy has been extended with minor modifications to Chinese (Wong and Ginzburg, 2013), French (Guida, 2013), Spanish (Garcia-Marchena, 2015), and Twitter (citation suppressed). Moreover, this taxonomy can be learnt using supervised (Fernández et al., 2007) and semi-supervised (Dragone and Lison, 2015) methods. Given that NSUs represent a wide

	NSU Class	Example		NSU Class	Example
1	Plain Acknowledgement (**Ack**)	*A:* ... *B:* mmh.	9	Propositional Modifier (**PropMod**)	*A:* Did Bo leave? *B:* Maybe.
2	Repeated Acknowledgement (**RepAck**)	*A:* Did Bo leave? *B:* Bo, hmm.	10	Rejection (**Reject**)	*A:* Did Bo leave? *B:* No.
3	Clarification Ellipsis (**CE**)	*A:* Did Bo leave? *B:* Bo?	11	Helpful Rejection (**HelpReject**)	*A:* Did Bo leave? *B:* No, Max.
4	Sluice (**Sluice**)	*A:* Someone left. *B:* Who?	12	Factive Modifier (**FactMod**)	*A:* Bo left. *B:* Great!
5	Check Question (**CheckQ**)	*A:* Bo isn't here. Okay?	13	Bare Modifier Phrase (**BareModPh**)	*A:* Max left. *B:* Yesterday.
6	Short Answer (**ShortAns**)	*A:* Who left? *B:* Bo.	14	Conjunction + Fragment (**Conj+Frag**)	*A:* Bo left. *B:* And Max.
7	Affirmative Answer (**AffAns**)	*A:* Did Bo leave? *B:* Yes.	15	Filler (**Filler**)	*A:* Did Bo ... *B:* leave?
8	Repeated Affirmative Answer (**RepAffAns**)	*A:* Did Bo leave? *B:* Bo, yes.			

Table 4: A Taxonomy for non-sentential utterances (NSUs)

variety of move types, one can hypothesize that **NSU distributions yield an "interactional profile" of a given conversational type.**

As a starting point for the current work, we investigated the frequency distribution of NSUs across the 23 BNC spoken genres. Files of total size in the range of 15,000-19,999 words were randomly selected from each genre, resulting in a sub-corpus consisting of 69 files, totalling 383,979 words. Annotation was manual, using the taxonomy of Fernández and Ginzburg (2002), the reliability of which is discussed in Fernández (2006). Table 5 shows the frequency distribution of NSUs across the 23 BNC spoken genres we obtained in that study, normalized here to 10,000 sentence units. As might be expected, those genres which are more interactive in nature (e.g., interview, medical consultation, classroom, and conversation) have high frequencies of NSUs, whereas those genres which are not interactive in nature (e.g., broadcast news, parliament, and sermon) have low frequencies of NSUs. On the basis of the data in Table 5, we calculated the probability distribution of the 15 NSU classes in each genre. The probability of occurrence of NSUs in a NSU class in a genre is the ratio of the normalized frequency of that NSU

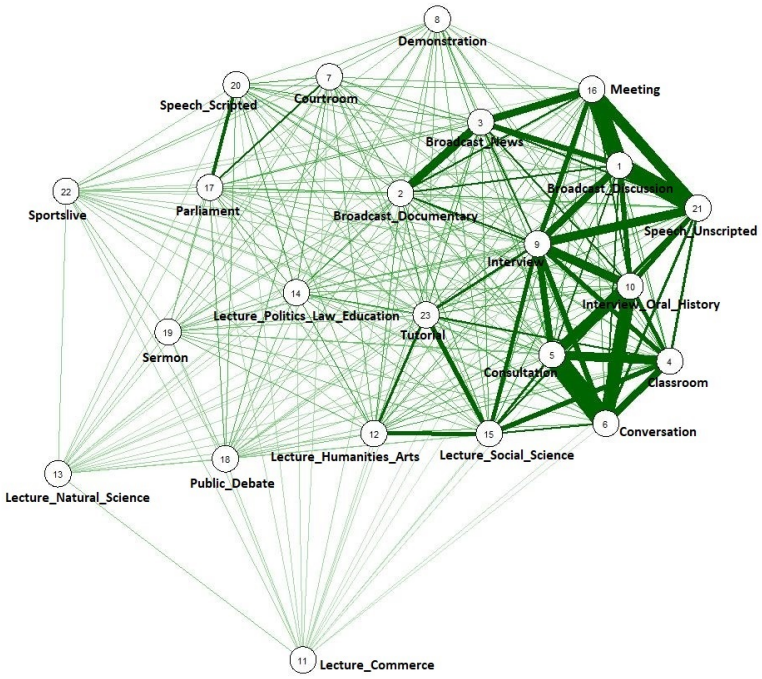

Figure 1: JSD metric of BNC spoken genres using unigrams

class in the genre by the total normalized frequency of the 15 NSU classes in the genre. These figures were used to generate the JSD metric among conversational types. Figure 3 displays this data using an FDG. The distance matrix for this metric sorted by closest neighbour is displayed in full in Table 12 in the Appendix.

Genre	Ack	RepAck	CE	Sluice	CheckQ	ShortAns	AffAns	RepAffAns	PropMod	Reject	HelpReject	FactMod	BareModPh	Conj+Frag	Filler	Total
1 Discn	1529	90	72	45	18	162	144	18	36	90	0	0	18	9	9	2240
2 Doc	62	10	41	21	0	21	10	0	0	21	0	10	0	0	0	196
3 News	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4 Class	1488	262	53	27	102	404	169	18	9	71	18	44	0	4	53	2722
5 Cons	1893	69	110	8	57	57	297	46	11	126	27	27	8	0	42	2778
6 Conv	1070	79	360	67	40	171	454	18	15	171	24	82	6	3	12	2572
7 Court	1010	57	38	0	0	114	133	19	38	105	10	19	0	0	0	1543
8 Demo	941	112	11	22	56	549	258	45	11	146	0	22	0	11	90	2274
9 Intv	2053	77	55	0	133	11	144	11	44	28	0	55	0	11	33	2655
10 Hist	2552	183	67	0	18	79	183	37	6	67	24	37	24	37	49	3363
11 Comm	74	25	0	0	0	0	25	0	0	0	0	0	0	0	0	124
12 H_arts	1058	50	40	10	0	40	190	0	20	20	20	0	0	0	0	1448
13 Nat_sc	157	14	29	0	157	143	86	0	0	0	0	14	0	0	0	600
14 P_law	78	155	58	0	0	388	19	19	0	0	19	78	0	0	0	814
15 Soc_sc	233	78	13	13	0	39	65	0	0	26	13	13	0	0	13	506
16 Meet	1024	70	42	7	49	63	181	14	28	42	0	14	0	7	42	1583
17 Prlmnt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18 P_deb	888	9	28	0	0	28	227	19	28	57	9	19	0	0	19	1331
19 Sermn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20 Sp_s	313	84	42	0	0	21	94	0	0	31	10	0	0	10	0	605
21 Sp_us	519	161	66	0	22	278	95	22	22	51	0	44	22	0	15	1317
22 Sport	78	34	0	0	0	9	0	9	0	9	9	0	9	9	9	175
23 Tut	916	44	71	0	0	62	169	53	9	44	9	36	0	0	18	1431

Table 5: Frequency distribution of NSUs across BNC spoken genres

4 Evaluation

How to compare the different metrics on the space of conversational types? We will do so by inspecting the neighbourhoods (k-nearest neighbours) of a given conversational type and consider the plausibility and robustness of the assigned neighbourhoods. *A priori* the situation is somewhat tricky—we have no

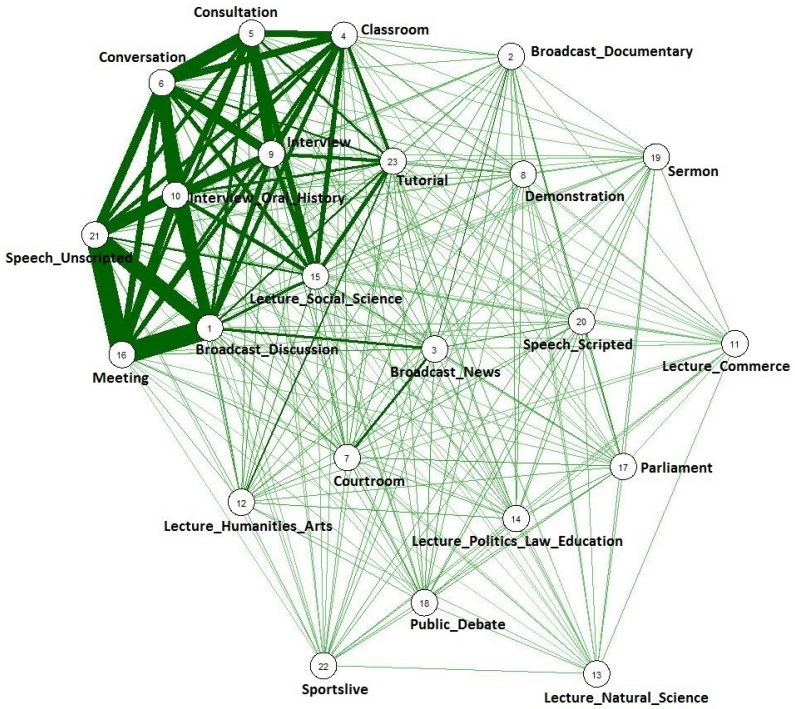

Figure 2: JSD metric of BNC spoken genres using bigrams

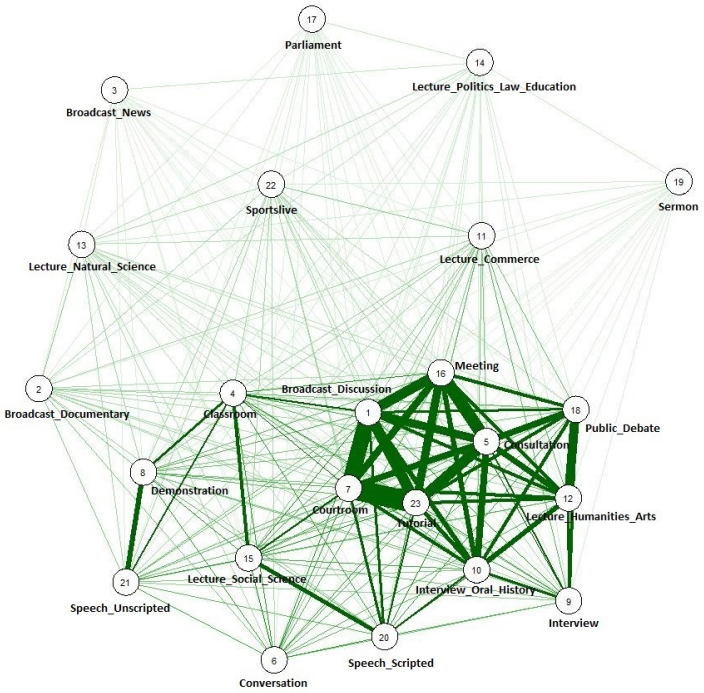

Figure 3: JSD metric of BNC spoken genres using NSUs

inconvertible gold standard to guide us. Nonetheless, we can propose some basic constraints, which allow us to compare the different metrics which take into account notions of interactivity and subject matter.

No-Interaction types Examining the class of conversational types, we can recognize three where (essentially) no interaction takes place: the classes concerned are *Broadcast News*(3), *Parliament*(17), and *Sermon*(19). The defining principle of such types that can be summarized for an agent who needs to be taught how to participate is that the agent is in such cases an *overhearer* (Goffman, 1981), who does not speak. ("Don't speak back to the tv or during a sermon/speech.") The lack of interactivity is captured well by their null NSU distributions. This means that the NSU-based metric isolates these types as a cluster. On the other hand, the uni/bi-gram-based methods do not capture this requirement, yielding the following neighbourhoods (extracted from Tables 10, 11, 12 in the Appendix):[4]

Genre	Nearest Neighbours	Method	Genre	Nearest Neigbhours	Method
3	1,16,[2,21],10,20	Unigram	19	[6,1],[21,10],4,[15,16,12, 5,9],3	Unigram
	1,16,21,10,20	Bigram		[21,10,1],16,6,4,[9,15]	Bigram
	17,**19**	NSU		**3**,**17**	NSU
17	16,3,1,20,[7,21,2]	Unigram			
	16,3,1,21,7	Bigram			
	3,**19**	NSU			

Table 6: Nearest 5 neighbours for non-interactive types

Types with similar subject matter: difference As we noted in the introduction, the guiding principle of current formal models for conversational types is largely driven by subject matter. Thus, a fixed set of questions (via domain issues, QNUD etc) is essentially a defining characteristic of a conversational type. This is problematic in two ways. For a start, types in principle can share subject matter but differ because of distinct interactional organization. In the BNC collection of types this is exemplified by types *Parliament*(17) and *Public Debate*(18). The NSU metric isolates these two types from each other and, intuitively, places *Public Debate*(18) closest to various 'uncontrolled interaction types' such as *Meeting*(16), *Consultation*(5), and *Interview*(9); the uni/bi-gram metrics, not surprisingly place the two types among their closest neighbours.

Genre	Nearest Neighbours	Method	Genre	Nearest Neigbhours	Method
17	16,3,1,20,[7,21,2],14,9,23,**18**	Unigram	18	16,1,[3,21],**17**,7,23,[2,20,9],14,[4,15]	Unigram
	16,3,1,21,7,20,[10,2],**18**,9	Bigram		16,7,1,21,[3,10],[6,4],23,[9,**17**],5	Bigram
	3,19	NSU		[23,5],[7,12,16],10,1,9,6,20,[15,4],8	NSU

Table 7: Nearest 9 neighbours for types concerning parliament

Complex subject matter structure: Sportslive Another problem for methods based on a simple characterization of subject matter is a type like *Sportslive*(22) (commentary), which involves a main commentator exchanging impressions on an ongoing sports event with an additional (expert/side) commentator. This type has low but non-zero NSU frequency (Ack: 78, Repack: 34, ShortAns: 9, RepAffAns: 9, Reject: 9, HelpReject: 9, BareModPh: 9, Conj+Frag: 9, Filler: 9) and essentially involves a repeated question: *what's going on now?* (along with issues raised by answers to the different tokens of this question). The NSU-based method, as with the type *Public Debate*(18) discussed above, places *Sportslive*(22) (commentary) closest to various 'uncontrolled interaction types'; the uni/bi-gram-based methods do, on the whole, well on this type too, locating it next to types such as *Classroom*(4) and (medical) *Consultation*(5). However, they also place it next to the non-interactive type *Broadcast News*(3):

Genre	Nearest Neighbours	Method
22	1,6,21,[5,10],4,[16,**3**]	Unigram
	1,21,[16,**3**,6],10,4,[5,9]	Bigram
	10,[15,4],[20,21],1,7,[23,16,5,8]	NSU

Table 8: Nearest 6 neighbours for the *Sportslive*(22) (commentary) type

[4]The notation [a,b,...] means that the types a,b,... all have the same distance from the given type.

Types with similar subject matter: similarity among the lecture types The NSU-based metric captures the apparent generalization that (apart from lecture type *Lecture Natural Science*(13), which by all methods seems to be somewhat distinct) all the lecture types, including *Lecture Commerce*(11), *Lecture Humanities Arts*(12), *Lecture Natural Science*(13), *Lecture Politics Law Education*(14), and *Lecture Social Science*(15), are close neighbours better than the uni/bi-gram-based metrics:

Genre	Nearest Neighbours	Method	Genre	Nearest Neigbhours	Method
11	23,21,4,16,**15**,9	Unigram	14	1,[21,16],[3,23],[**12**,17],2,[**15**,9]	Unigram
	4,21,23,[10,16],9,[1,**15**]	Bigram		21,1,16,10,3,[23,**15**]	Bigram
	20,**12**,[10,**15**],16,[18,5,23,9],7	NSU		21,[4,8],**13**,[**15**,2],6,20	NSU
12	**15**,1,23,21,[10,16],9	Unigram	15	21,1,**12**,[4,9],[16,23,5],6	Unigram
	10,23,[21,**15**],1,16,[6,9]	Bigram		21,5,[10,6,16],9,4,1	Bigram
	[18,7,5],[1,23,16,10],20,9,[**15**,4,**11**],6	NSU		[4,20],[23,6,7],[16,21,5,1],[8,**12**,10],18,**11**	NSU
13	4,21,**15**,23,1,16	Unigram			
	21,4,16,1,10,[23,**15**,6]	Bigram			
	4,[8,21],6,16,[**15**,5,1,23,7],9	NSU			

Table 9: Nearest 6 neighbours for lecture types

5 Discussion

Section 4 shows that for a variety of cases a metric based on NSU distributions imposes a more convincing topological structure on the class of conversational types than a metric based on uni/bi-grams.

This confirms our hypothesis from section 3 that this distribution constitutes an "interactional profile" of a conversational type. It provides us with a potential operational criterion when encountering a novel conversational domain—situating it within the class of conversational types can be achieved by sampling its NSUs and evaluating the emergent distribution relative to existing NSU distributions.

This has a significant implication for existing models of conversational types. These place the burden of variation among types in terms of subject matter and moves, while assuming that the conversational principles (e.g., the potential for either a grounding move or a clarification move as a follow up to any given move) are general. However, metrics based on such notions, as exemplified by uni/bi-gram-based metrics, are intrinsically too coarse. The consequence is that the specification of conversational types must also include the specification of distinct *neighbourhoods*, collections of similar types, governed by conversational principles that apply specifically to them (e.g., one class of types enables clarification interaction to be triggered at turn exchange junctures, whereas in others such a potential does not exist.).

6 Conclusions and Future Work

The notion of a conversational type (aka *language game, speech/conversational genre*) originates in philosophy of language and pragmatics. It is one of the fundamental notions of dialogue, embodying those aspects that serve to characterize domain specific aspects of interaction, both in terms of relevance and choice of forms. There exist theoretical models of this notion, but attempts at global characterization of the space of types and specifically defining (distance) metrics for the entire space has not, as far as we are aware, been attempted before.

We use both uni/bi-gram-based metrics and a metric based on the distribution of non-sentential utterances (NSUs). We argue for the superiority of metrics based on non-sentential utterance distributions, though the uni/bi-gram-based metrics also yield plausible results.

Although we have related given 'atomic' types, based on the BNC taxonomy, our method does not depend on this and we could in future work apply this approach to a corpus without predefining partitions. We have used the BNC, given the wide range of types it contains. But it is of course important to investigate such metrics using balanced corpora in other languages (e.g., the Swedish Gothenburg corpus (Allwood, 1999) and the Polish National Corpus (Przepiórkowski et al., 2008).). We also plan to refine the NSU-based metric to include additional interactional features such as disfluencies or laughter, which vary significantly across conversational types (Hough et al., 2016).

From a theoretical point of view, we have argued that the results of our experiments force one to rethink the notion of conversational type to incorporate aspects that go beyond subject matter and form, by incorporating, for instance, parameters that relate to turn control and participant autonomy.

Acknowledgements

This research was supported by the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083) and a senior fellowship to the second author by the Institut Universitaire de France. In addition, we would like to thank three anonymous reviewers for SemDial for their useful comments.

References

Jens Allwood. 1995. An activity based approach of pragmatics. Technical report, Gothenburg Papers in Theoretical Linguistics, 76. Reprinted in Bunt et al (2000) 'Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics'. Amsterdam, John Benjamins.

Jens Allwood. 1999. The swedish spoken language corpus at göteborg university. In *Proceedings of Fonetik 99*, volume 81 of *Gothenburg Papers in Theoretical Linguistics*.

L. Anthony. 2017. Antconc [computer software]. Version: 3.5.2.

M.M. Bakhtin. 1986. *Speech Genres and Other Late Essays*. University of Texas Press.

Michael J. Bannister, David Eppstein, Michael T. Goodrich, and Lowell Trott. 2012. Force-directed graph drawing using social gravity and scaling. *CoRR*, abs/1209.0748.

Douglas Biber and Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2):95–137.

L. Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford Universtity Computing Services.

Stephen Clark. 2015. Vector space models of lexical meaning. *Handbook of Contemporary Semantic Theory, The*, pages 493–522.

Philip Cohen and Ray Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:177–212.

Paolo Dragone and Pierre Lison. 2015. An active learning approach to the classification of non-sentential utterances. *CLiC it*, page 115.

Raquel Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College, London.

Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances: A corpus study. *Traitement automatique des languages. Dialogue*, 43(2):13–42.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying ellipsis in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, page 31. IEEE.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Erving Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press, Philadelphia.

Flore Guida. 2013. Les phrases sans verbes. Université Paris- Diderot Ms.

S. Hoffmann, S. Evert, N. Smith, D. Lee, and Y. Berglund-Prytz. 2008. *Corpus linguistics with BNCweb-a practical guide (Vol. 6)*. Peter Lang.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *Proceedings of LREC 2016*.

Dell Hymes. 1972. On communicative competence. *sociolinguistics*, 269:293.

Staffan Larsson. 2002. *Issue based Dialogue Management*. Ph.D. thesis, Gothenburg University.

Staffan Larsson and Alexander Berman. 2016. Domain-specific and general syntax and semantics in the talkamatic dialogue manager. *Empirical Issues in Syntax and Semantics*, 11:91–110.

Stephen C Levinson. 1979. Activity types and language. *Linguistics*, 17(5-6):365–400.

Diane Litman and James Allen. 1987. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200.

C. Manning and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, Mass.

Adam Przepiórkowski, Rafal L Górski, Barbara Lewandowska-Tomaszyk, and Marek Lazinski. 2008. Towards the national corpus of polish. In *LREC*.

R-Core-Team. 2013. R: A language and environment for statistical computing.

Roger C Schank and Robert Abelson. 1977. *Scripts, goals, plans, and understanding*. Hillsdale, NJ: Erlbaum.

David Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, University of Edinburgh, Edinburgh.

C. M. Tan, Y. F. Wang, and C. D. Lee. 2002. The use of bigrams to enhance text categorization. *Information processing and management*, 38(4):529–546.

I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford. Citations from second edition, 1988 reprint.

Kwong-Cheong Wong and Jonathan Ginzburg. 2013. Investigating non-sentential utterances in a spoken chinese corpus. In *PACLING 2013*.

Appendix

	1	Discn	2	Doc	3	News	4	Class	5	Cons	6	Conv	7	Court	8	Demo	9	Intv	10	Hist	11	Comm	12	H_arts	13	Nat_sc	14	P_law	15	Soc_sc	16	Meet	17	Prlmnt	18	P_deb	19	Sermn	20	Sp_s	21	Sp_us	22	Sport	23	Tut
1st	16	0.08	3	0.15	1	0.09	21	0.13	6	0.10	10	0.10	16	0.21	5	0.20	21	0.09	6	0.10	23	0.42	15	0.17	4	0.39	1	0.22	21	0.15	21	0.08	16	0.19	16	0.24	6	0.30	16	0.19	16	0.08	1	0.32	21	0.17
2nd	21	0.08	1	0.15	16	0.11	6	0.14	21	0.11	5	0.10	21	0.24	6	0.20	1	0.13	21	0.11	21	0.45	1	0.19	21	0.41	21	0.24	1	0.16	1	0.08	3	0.22	1	0.30	1	0.30	3	0.19	1	0.08	6	0.34	1	0.18
3rd	3	0.09	21	0.17	2	0.15	5	0.15	10	0.13	21	0.11	1	0.25	4	0.23	16	0.14	1	0.12	4	0.47	23	0.21	15	0.42	16	0.24	12	0.17	3	0.11	1	0.23	3	0.32	21	0.31	1	0.23	9	0.09	21	0.35	16	0.19
4th	10	0.12	16	0.17	21	0.15	1	0.17	4	0.15	4	0.14	3	0.26	21	0.23	6	0.15	5	0.13	16	0.48	21	0.23	23	0.44	3	0.25	4	0.19	9	0.14	20	0.24	21	0.32	10	0.31	17	0.24	6	0.11	5	0.36	15	0.20
5th	9	0.13	9	0.22	10	0.18	15	0.19	9	0.16	1	0.15	17	0.26	10	0.26	10	0.15	9	0.15	15	0.49	10	0.25	1	0.45	23	0.25	9	0.19	10	0.16	7	0.26	17	0.33	4	0.33	21	0.25	5	0.11	10	0.36	12	0.21
6th	6	0.15	10	0.23	20	0.19	16	0.19	1	0.16	9	0.15	2	0.28	9	0.28	5	0.16	16	0.16	9	0.50	16	0.25	16	0.46	12	0.27	16	0.20	6	0.17	21	0.26	7	0.34	15	0.34	2	0.26	10	0.11	4	0.37	9	0.22
7th	2	0.15	20	0.26	9	0.21	10	0.19	16	0.17	16	0.17	9	0.28	1	0.29	15	0.19	3	0.18	1	0.51	9	0.26	9	0.46	17	0.27	23	0.20	5	0.17	2	0.26	23	0.35	16	0.34	14	0.30	4	0.13	16	0.38	4	0.23
8th	15	0.16	23	0.26	17	0.22	9	0.19	8	0.20	8	0.20	10	0.31	16	0.30	4	0.19	4	0.19	14	0.51	14	0.27	5	0.47	2	0.28	5	0.20	2	0.17	14	0.27	2	0.38	12	0.34	9	0.31	3	0.15	3	0.38	3	0.24
9th	5	0.16	17	0.26	6	0.24	8	0.23	15	0.20	15	0.21	5	0.32	15	0.32	3	0.21	15	0.22	5	0.51	3	0.29	14	0.48	15	0.29	6	0.21	17	0.19	9	0.31	20	0.38	5	0.34	23	0.31	15	0.15	9	0.39	5	0.24
10th	4	0.17	5	0.26	23	0.24	23	0.23	23	0.24	3	0.24	6	0.33	23	0.35	23	0.22	2	0.23	3	0.52	4	0.29	3	0.49	9	0.29	10	0.22	4	0.19	23	0.32	9	0.38	9	0.34	10	0.31	23	0.17	8	0.41	10	0.24
11th	23	0.18	6	0.27	14	0.25	3	0.26	3	0.25	23	0.25	14	0.34	3	0.37	2	0.22	23	0.24	2	0.53	5	0.30	2	0.50	20	0.30	2	0.27	20	0.19	18	0.33	14	0.39	3	0.36	7	0.35	2	0.17	2	0.42	6	0.25
12th	12	0.19	15	0.27	5	0.25	2	0.28	2	0.26	2	0.27	18	0.34	2	0.39	12	0.26	12	0.25	13	0.54	6	0.30	10	0.50	10	0.31	3	0.28	23	0.19	10	0.34	4	0.40	2	0.37	4	0.35	12	0.23	15	0.43	14	0.25
13th	14	0.22	7	0.28	7	0.26	12	0.29	12	0.30	12	0.30	23	0.34	12	0.40	8	0.28	8	0.26	6	0.54	2	0.31	12	0.50	4	0.32	14	0.29	15	0.20	12	0.36	15	0.40	23	0.37	5	0.36	8	0.23	23	0.43	2	0.26
14th	20	0.23	14	0.28	4	0.26	14	0.32	7	0.32	19	0.30	4	0.35	19	0.40	7	0.28	19	0.31	12	0.56	19	0.34	6	0.51	7	0.34	8	0.32	7	0.21	15	0.37	10	0.41	14	0.40	6	0.36	14	0.24	12	0.46	20	0.31
15th	17	0.23	4	0.28	15	0.28	19	0.33	19	0.34	7	0.33	20	0.35	22	0.41	14	0.29	7	0.31	10	0.56	17	0.36	20	0.52	5	0.34	19	0.34	18	0.24	4	0.38	5	0.42	8	0.40	15	0.37	7	0.24	19	0.48	17	0.32
16th	7	0.25	12	0.31	12	0.29	7	0.35	14	0.34	22	0.34	15	0.35	14	0.42	20	0.31	14	0.31	20	0.56	7	0.38	8	0.52	6	0.35	7	0.35	14	0.24	5	0.38	12	0.42	7	0.41	12	0.38	20	0.25	7	0.48	7	0.34
17th	8	0.29	19	0.37	18	0.32	20	0.35	22	0.36	14	0.35	12	0.38	7	0.44	17	0.31	20	0.31	7	0.56	20	0.38	7	0.53	18	0.39	17	0.37	12	0.25	6	0.39	6	0.43	17	0.45	18	0.38	17	0.26	14	0.49	8	0.35
18th	19	0.30	18	0.38	19	0.36	22	0.37	20	0.36	20	0.36	19	0.41	20	0.46	19	0.34	17	0.34	8	0.57	8	0.40	18	0.53	19	0.40	20	0.37	8	0.30	19	0.45	8	0.50	20	0.45	19	0.45	19	0.31	17	0.52	18	0.35
19th	18	0.30	8	0.39	8	0.37	17	0.38	17	0.38	17	0.39	8	0.44	17	0.49	18	0.38	22	0.36	18	0.58	18	0.42	17	0.53	8	0.42	18	0.40	19	0.34	8	0.49	19	0.51	22	0.48	8	0.46	18	0.32	20	0.52	19	0.37
20th	22	0.32	22	0.42	22	0.38	13	0.39	18	0.42	18	0.43	22	0.48	18	0.50	22	0.39	18	0.41	17	0.58	22	0.46	11	0.54	13	0.48	13	0.42	22	0.38	22	0.52	22	0.53	18	0.51	13	0.52	22	0.35	18	0.53	11	0.42
21st	13	0.45	13	0.50	13	0.49	18	0.40	13	0.47	13	0.51	13	0.53	13	0.52	13	0.46	13	0.50	19	0.63	13	0.50	19	0.58	22	0.49	22	0.43	13	0.46	13	0.53	13	0.53	13	0.58	22	0.52	13	0.41	13	0.60	22	0.43
22nd	11	0.51	11	0.53	11	0.52	11	0.47	11	0.51	11	0.54	11	0.56	11	0.57	11	0.50	11	0.56	22	0.64	11	0.56	22	0.60	11	0.51	11	0.49	11	0.48	11	0.58	11	0.58	11	0.63	11	0.56	11	0.45	11	0.64	13	0.44

Table 10: Nearest neighbours among BNC spoken genres using unigrams

	1	Discn	2	Doc	3	News	4	Class	5	Cons	6	Conv	7	Court	8	Demo	9	Intv	10	Hist	11	Comm	12	H_arts	13	Nat_sc	14	P_law	15	Soc_sc	16	Meet	17	Prlmnt	18	P_deb	19	Sermn	20	Sp_s	21	Sp_us	22	Sport	23	Tut
1st	16	0.21	16	0.51	1	0.29	21	0.30	6	0.25	21	0.23	16	0.40	21	0.50	21	0.27	6	0.23	4	0.72	10	0.47	21	0.75	21	0.60	21	0.34	21	0.17	16	0.53	16	0.50	21	0.65	3	0.51	16	0.17	1	0.59	21	0.38
2nd	21	0.22	21	0.52	16	0.37	6	0.31	21	0.29	10	0.23	21	0.41	6	0.50	5	0.30	21	0.24	21	0.74	23	0.49	4	0.76	1	0.61	5	0.36	1	0.21	3	0.57	7	0.56	10	0.65	16	0.51	1	0.22	21	0.65	16	0.41
3rd	10	0.28	3	0.53	21	0.42	16	0.34	9	0.30	5	0.25	1	0.46	5	0.52	10	0.32	16	0.26	23	0.75	21	0.50	16	0.78	16	0.64	10	0.38	6	0.25	1	0.59	1	0.57	1	0.65	1	0.53	6	0.23	16	0.67	10	0.42
4th	3	0.29	1	0.53	10	0.50	5	0.37	10	0.30	16	0.25	10	0.50	10	0.54	16	0.32	1	0.28	10	0.76	15	0.50	1	0.79	10	0.65	6	0.38	10	0.26	21	0.60	21	0.58	16	0.68	21	0.56	10	0.24	3	0.67	15	0.44
5th	6	0.31	10	0.59	20	0.51	9	0.40	16	0.31	4	0.31	3	0.52	4	0.55	6	0.32	5	0.30	16	0.76	1	0.51	10	0.80	3	0.66	16	0.38	5	0.31	7	0.61	3	0.64	6	0.69	7	0.63	9	0.27	6	0.67	5	0.45
6th	9	0.36	7	0.62	7	0.52	1	0.41	15	0.36	1	0.31	6	0.55	9	0.56	1	0.36	9	0.32	9	0.77	16	0.53	23	0.81	23	0.67	9	0.41	9	0.32	20	0.65	10	0.64	4	0.72	10	0.65	5	0.29	10	0.69	9	0.45
7th	5	0.38	9	0.62	2	0.53	15	0.42	4	0.37	9	0.32	18	0.56	16	0.57	4	0.40	15	0.38	1	0.78	6	0.57	15	0.81	15	0.67	4	0.42	4	0.34	10	0.66	6	0.68	9	0.73	17	0.65	4	0.30	4	0.72	1	0.45
8th	4	0.41	6	0.63	6	0.53	10	0.43	1	0.38	15	0.38	9	0.57	15	0.58	15	0.41	23	0.42	15	0.78	9	0.57	6	0.81	9	0.68	1	0.43	3	0.37	2	0.66	4	0.68	15	0.73	4	0.66	15	0.34	5	0.73	6	0.45
9th	15	0.43	5	0.66	17	0.57	23	0.45	23	0.45	23	0.45	4	0.58	1	0.59	23	0.45	4	0.43	6	0.79	5	0.59	5	0.82	12	0.70	23	0.44	15	0.38	18	0.70	23	0.69	3	0.74	6	0.66	23	0.38	9	0.73	4	0.45
10th	23	0.45	17	0.66	9	0.58	8	0.55	8	0.52	8	0.50	5	0.60	23	0.62	8	0.56	12	0.47	5	0.79	4	0.61	9	0.82	4	0.71	12	0.50	7	0.40	9	0.72	9	0.70	12	0.74	2	0.68	7	0.41	15	0.75	12	0.49
11th	7	0.46	15	0.67	4	0.58	7	0.58	12	0.59	3	0.53	17	0.61	12	0.70	7	0.57	3	0.50	8	0.82	3	0.68	7	0.83	5	0.71	8	0.58	23	0.41	23	0.73	17	0.70	23	0.74	9	0.70	3	0.42	7	0.76	7	0.61
12th	12	0.51	23	0.67	5	0.62	3	0.58	7	0.60	7	0.55	23	0.61	3	0.75	12	0.57	7	0.50	3	0.82	7	0.69	3	0.84	6	0.71	7	0.65	18	0.50	6	0.74	5	0.71	5	0.74	18	0.72	12	0.50	8	0.77	8	0.62
13th	20	0.53	4	0.67	18	0.64	12	0.61	3	0.62	12	0.57	2	0.62	7	0.76	3	0.58	8	0.54	12	0.83	14	0.70	8	0.85	7	0.72	3	0.66	2	0.51	14	0.75	20	0.72	8	0.78	23	0.72	8	0.50	12	0.78	3	0.64
14th	2	0.53	20	0.68	23	0.64	20	0.66	2	0.66	2	0.63	20	0.63	22	0.77	2	0.62	2	0.59	14	0.83	8	0.70	12	0.85	17	0.75	2	0.67	20	0.51	4	0.76	15	0.72	7	0.78	5	0.74	2	0.52	2	0.78	2	0.67
15th	18	0.57	12	0.72	15	0.66	2	0.67	18	0.71	20	0.66	15	0.65	2	0.78	14	0.68	18	0.64	7	0.85	2	0.72	2	0.85	2	0.75	14	0.67	17	0.53	15	0.77	2	0.74	2	0.79	15	0.75	20	0.56	23	0.79	14	0.67
16th	8	0.59	18	0.74	14	0.66	18	0.68	14	0.71	22	0.67	12	0.69	19	0.78	18	0.70	14	0.65	2	0.85	19	0.74	18	0.86	20	0.77	18	0.72	12	0.53	5	0.77	12	0.76	14	0.80	12	0.76	18	0.58	20	0.80	18	0.69
17th	22	0.59	14	0.75	22	0.67	14	0.71	22	0.73	18	0.68	14	0.72	14	0.80	20	0.70	20	0.65	20	0.85	18	0.76	14	0.86	18	0.79	19	0.73	8	0.57	12	0.78	14	0.79	22	0.81	14	0.77	14	0.60	19	0.81	20	0.72
18th	17	0.59	8	0.78	12	0.68	11	0.72	20	0.74	19	0.69	22	0.76	11	0.82	17	0.72	19	0.65	18	0.87	20	0.76	20	0.86	8	0.80	20	0.75	14	0.64	22	0.83	8	0.82	20	0.82	22	0.80	17	0.60	18	0.83	17	0.73
19th	14	0.61	22	0.78	19	0.74	22	0.72	19	0.74	14	0.71	8	0.76	18	0.82	19	0.73	17	0.66	19	0.89	17	0.78	17	0.87	19	0.80	22	0.75	22	0.67	19	0.83	22	0.83	17	0.83	19	0.82	22	0.65	17	0.83	19	0.74
20th	19	0.65	19	0.79	8	0.75	19	0.72	17	0.77	17	0.74	19	0.78	20	0.83	22	0.73	22	0.69	22	0.89	22	0.78	19	0.87	11	0.83	17	0.77	19	0.68	8	0.87	19	0.84	18	0.84	8	0.83	19	0.65	14	0.85	11	0.75
21st	11	0.78	11	0.85	11	0.82	13	0.76	11	0.79	11	0.79	13	0.83	13	0.85	11	0.77	11	0.76	17	0.90	11	0.83	22	0.90	22	0.85	11	0.78	11	0.76	13	0.87	13	0.86	13	0.87	11	0.85	11	0.74	11	0.89	22	0.79
22nd	13	0.79	13	0.85	13	0.84	17	0.76	13	0.82	13	0.81	11	0.85	17	0.87	13	0.82	13	0.80	13	0.91	13	0.85	11	0.91	13	0.86	13	0.81	13	0.78	11	0.90	11	0.87	11	0.89	13	0.86	13	0.75	13	0.90	13	0.81

Table 11: Nearest neighbours among BNC spoken genres using bigrams

	1	Discn	2	Doc	3	News	4	Class	5	Cons	6	Conv	7	Court	8	Demo	9	Intv	10	Hist	11	Comm	12	H_arts	13	Nat_sc	14	P_law	15	Soc_sc	16	Meet	17	Prlmnt	18	P_deb	19	Sermn	20	Sp_s	21	Sp_us	22	Sport	23	Tut
1st	16	0.03	6	0.09	17	0.00	8	0.05	16	0.02	23	0.08	23	0.03	4	0.05	16	0.04	5	0.03	20	0.10	18	0.04	4	0.18	21	0.20	4	0.06	5	0.02	3	0.00	23	0.03	3	0.00	15	0.06	4	0.05	10	0.17	5	0.02
2nd	7	0.04	15	0.16	19	0.00	21	0.05	23	0.02	5	0.08	1	0.04	21	0.06	10	0.04	16	0.04	12	0.11	7	0.04	8	0.19	4	0.30	20	0.06	1	0.03	19	0.00	5	0.03	17	0.00	12	0.07	8	0.06	15	0.19	18	0.03
3rd	5	0.04	21	0.17	4	0.50	16	0.05	18	0.03	2	0.09	18	0.04	16	0.09	5	0.05	9	0.04	10	0.14	5	0.04	21	0.19	8	0.30	23	0.09	9	0.04	4	0.50	7	0.04	4	0.50	7	0.08	15	0.10	4	0.19	7	0.03
4th	10	0.04	1	0.19	5	0.50	1	0.06	10	0.03	16	0.09	5	0.04	1	0.10	1	0.08	23	0.04	15	0.14	1	0.05	6	0.23	13	0.32	6	0.09	10	0.04	5	0.50	12	0.04	5	0.50	23	0.09	7	0.10	20	0.21	16	0.04
5th	12	0.05	4	0.19	6	0.50	15	0.06	1	0.04	15	0.09	12	0.04	15	0.11	12	0.08	1	0.04	16	0.15	23	0.05	16	0.25	15	0.33	7	0.09	23	0.04	6	0.50	16	0.04	6	0.50	5	0.09	1	0.10	21	0.21	10	0.04
6th	23	0.06	20	0.20	10	0.50	7	0.07	7	0.04	7	0.10	16	0.05	7	0.11	18	0.08	12	0.05	18	0.16	16	0.05	15	0.29	2	0.33	16	0.10	18	0.04	10	0.50	10	0.06	10	0.50	10	0.09	6	0.11	1	0.22	12	0.05
7th	4	0.06	7	0.20	1	0.50	10	0.08	12	0.04	20	0.10	10	0.05	23	0.13	23	0.08	7	0.05	5	0.16	10	0.05	5	0.29	6	0.39	21	0.10	7	0.05	1	0.50	1	0.07	1	0.50	16	0.10	16	0.11	7	0.23	1	0.06
8th	18	0.07	23	0.20	8	0.50	5	0.08	9	0.05	1	0.10	4	0.07	6	0.13	7	0.09	18	0.06	23	0.16	20	0.07	1	0.29	20	0.42	5	0.10	4	0.05	8	0.50	9	0.08	8	0.50	1	0.10	23	0.12	23	0.24	6	0.08
9th	9	0.08	8	0.22	9	0.50	23	0.08	4	0.08	18	0.10	20	0.08	5	0.13	4	0.11	4	0.08	9	0.16	9	0.08	23	0.29	7	0.43	1	0.10	12	0.05	9	0.50	6	0.10	9	0.50	6	0.10	5	0.14	16	0.24	9	0.08
10th	20	0.10	5	0.23	16	0.50	9	0.11	6	0.08	4	0.11	9	0.09	18	0.15	20	0.16	20	0.09	7	0.17	15	0.11	7	0.29	23	0.43	8	0.11	6	0.09	16	0.50	20	0.11	16	0.50	11	0.10	20	0.14	5	0.24	4	0.08
11th	6	0.10	16	0.24	21	0.50	6	0.11	20	0.09	21	0.11	15	0.09	10	0.16	11	0.16	15	0.11	1	0.18	4	0.11	9	0.30	22	0.44	12	0.11	8	0.09	21	0.50	15	0.13	21	0.50	4	0.11	10	0.14	8	0.24	20	0.09
12th	8	0.10	12	0.26	7	0.50	12	0.11	15	0.10	12	0.12	6	0.10	20	0.17	6	0.17	11	0.14	4	0.19	11	0.11	20	0.31	1	0.46	10	0.11	15	0.10	7	0.50	4	0.13	7	0.50	18	0.11	2	0.17	12	0.28	15	0.09
13th	15	0.10	10	0.27	15	0.50	20	0.11	8	0.13	8	0.13	21	0.10	12	0.18	15	0.18	6	0.14	21	0.25	6	0.12	14	0.32	16	0.49	18	0.13	20	0.10	15	0.50	8	0.15	15	0.50	21	0.14	18	0.18	9	0.29	21	0.12
14th	21	0.10	18	0.27	18	0.50	18	0.13	21	0.14	10	0.14	8	0.11	13	0.19	21	0.19	21	0.14	6	0.25	8	0.18	2	0.32	10	0.50	11	0.14	21	0.11	18	0.50	11	0.16	18	0.50	9	0.16	12	0.18	11	0.29	8	0.13
15th	11	0.18	9	0.32	23	0.50	13	0.18	11	0.16	9	0.17	11	0.17	9	0.20	8	0.20	8	0.16	8	0.27	21	0.18	12	0.32	3	0.50	2	0.16	11	0.15	23	0.50	21	0.18	23	0.50	8	0.17	13	0.19	18	0.30	11	0.16
16th	2	0.19	13	0.32	2	0.50	11	0.19	2	0.23	13	0.23	2	0.20	2	0.22	22	0.29	22	0.17	22	0.29	2	0.26	18	0.33	17	0.50	9	0.18	2	0.24	2	0.50	2	0.27	2	0.50	2	0.20	9	0.19	6	0.33	2	0.20
17th	22	0.22	14	0.33	12	0.50	22	0.19	22	0.24	11	0.25	22	0.23	22	0.24	13	0.30	2	0.27	2	0.39	22	0.28	10	0.34	19	0.50	22	0.19	22	0.24	12	0.50	22	0.30	12	0.50	22	0.21	14	0.20	2	0.39	22	0.24
18th	13	0.29	11	0.39	14	0.50	2	0.19	13	0.29	22	0.33	13	0.29	11	0.27	2	0.32	13	0.34	13	0.40	13	0.32	11	0.40	5	0.52	13	0.29	13	0.25	14	0.50	13	0.33	14	0.50	13	0.31	22	0.21	14	0.44	13	0.29
19th	14	0.46	22	0.39	20	0.50	14	0.30	3	0.50	14	0.39	14	0.43	14	0.30	3	0.50	14	0.50	3	0.50	3	0.50	3	0.50	12	0.53	14	0.33	14	0.49	20	0.50	3	0.50	20	0.50	14	0.42	11	0.25	3	0.50	14	0.43
20th	3	0.50	3	0.50	22	0.50	3	0.50	17	0.50	3	0.50	3	0.50	3	0.50	17	0.50	3	0.50	17	0.50	17	0.50	17	0.50	11	0.55	3	0.50	3	0.50	22	0.50	17	0.50	22	0.50	3	0.50	3	0.50	17	0.50	3	0.50
21st	17	0.50	17	0.50	13	0.50	17	0.50	19	0.50	17	0.50	17	0.50	17	0.50	19	0.50	17	0.50	19	0.50	19	0.50	19	0.50	18	0.56	17	0.50	17	0.50	13	0.50	19	0.50	13	0.50	17	0.50	17	0.50	19	0.50	17	0.50
22nd	19	0.50	19	0.50	11	0.50	19	0.50	14	0.52	19	0.50	19	0.50	19	0.50	14	0.59	19	0.50	14	0.55	14	0.53	22	0.51	9	0.59	19	0.50	19	0.50	11	0.50	14	0.56	11	0.50	19	0.50	19	0.50	13	0.51	19	0.50

Table 12: Nearest neighbours among BNC spoken genres using NSUs

i	me	my	myself	we	our	won't	wouldn't	shan't	shouldn't	can't	cannot
ours	ourselves	you	your	yours	yourself	couldn't	mustn't	let's	that's	who's	what's
yourselves	he	him	his	himself	she	here's	there's	when's	where's	why's	how's
her	hers	herself	it	its	itself	a	an	the	and	but	if
they	them	their	theirs	themselves	what	or	because	as	until	while	of
which	who	whom	this	that	these	at	by	for	with	about	against
those	am	is	are	was	were	between	into	through	during	before	after
be	been	being	have	has	had	above	below	to	from	up	down
having	do	does	did	doing	would	in	out	on	off	over	under
should	could	ought	i'm	you're	he's	again	further	then	once	here	there
she's	it's	we're	they're	i've	you've	when	where	why	how	all	any
we've	they've	i'd	you'd	he'd	she'd	both	each	few	more	most	other
we'd	they'd	i'll	you'll	he'll	she'll	some	such	no	nor	not	only
we'll	they'll	isn't	aren't	wasn't	weren't	own	same	so	than	too	very
hasn't	haven't	hadn't	doesn't	don't	didn't						

Table 13: Stop words used in the experiments