Subjectivity of English discourse connectives
SUPPLEMENTARY MATERIAL

List of contents (S.I to S.IV)

I. Corpus analysis p. 2

II. Corpus sample analysis p. 2

1. Written material
    1. So
    2. Therefore
    3. As a result
    4. For this reason
2. Spoken material
    1. So
    2. Therefore
    3. As a result
    4. For this reason

III. Predictive analyses. Written data. p. 16

1. Relevance and expected sample sizes.
2. Crude probabilistic analysis.
3. Predictive analysis results, per domain, split by Rel and SoC.
4. Predictive analysis results, per domain, split by Modality

IV. Predictive analyses. Spoken data. p. 24

1. Relevance and expected sample sizes.
2. Crude predictive analysis
3. Predictive analysis results, per domain, split by Rel and SoC
4. Predictive analysis results, per domain, split by Modality

Section S.I. Corpus analysis

	Total connectives		Frequency per million words	Number or proportion of target instances
	Total written words 87,903,571	**Total spoken words 10,409,858**		
as a result/written	7584		86.28	1960 in total
as a result/spoken		387	37.18	70 in total
for this reason/written	964			789 in total
for this reason/spoken		15	1.44	6 in total
so/written	178216		2027.4	Examination of 2000 randomly selected sentences yielded 25% target instances. 250 of these 500 instances were used as target sample.
so/spoken		60900	5850.22	The sample of 250 target instances required 540 random sentences.
therefore/written	21406		243.52	The sample of 250 target instances required 276 random sentences.
therefore/spoken		1567	150.53	The sample of 250 target instances required 325 random sentences.

Table S1 Target instances of the analyzed connectives in the BNC.

Section S.II. Corpus sample analysis

S.II.1. Written material

1.1. So – domain summaries

> summary(SoWritten) All domains together

Rel	SoC	Modal	Domain sizes
SpA: 45	Au :64	0:209	Fiction : 75
Epi :106	CS:72	1: 11	NonAc :136
Vol : 59	Ch:32	2: 30	Newsp : 20
NonV : 40	Bl :42		Acad : 19
	(NonV : 40)		

> summary(SoWrittenF) Domain Fiction, size 75

Rel	SoC	Modal
SpA:12	Au :11	0:64
Epi :32	CS:33	1: 2
Vol :26	Ch:18	2: 9
NonV : 5	Bl : 8	
	(NonV : 5)	

> summary(SoWrittenNA) Domain NonAc, size 136

Rel	SoC	Modal
SpA:29	Au :40	0:114
Epi :55	CS:34	1: 8
Vol :27	Ch:11	2: 14
NonV :25	Bl :26	
	(NonV :25)	

> summary(SoWrittenN) Domain Newsp, size 20

Rel	SoC	Modal
SpA:3	Au :4	0:19
Epi:7	CS:5	1: 0
Vol :5	Ch:3	2: 1
NonV :5	Bl :3	
	(NonV :5)	

> summary(SoWrittenA) Domain Acad, size 19

Rel	SoC	Modal
SpA: 1	Au :9	0:12
Epi :12	CS:0	1: 1
Vol : 1	Ch:0	2: 6
NonV : 5	Bl :5	
	(NonV :5)	

1.1.1. So – Rel versus SoC

> table(SoWrittenF$Rel,SoWrittenF$Soc) Domain Fiction, size 75

	Au	CS	Ch	Bl	(NonV)
SpA	3	7	1	1	
Epi	8	13	4	7	
Vol	0	13	13	0	
NonV					5

> table(SoWrittenNA$Rel,SoWrittenNA$Soc) Domain NonAc, size 136

	Au	CS	Ch	Bl	NonV
SpA	16	3	1	9	
Epi	24	15	1	15	
Vol	0	16	9	2	
NonV					25

> table(SoWrittenN$Rel,SoWrittenN$Soc) Domain Newsp, size 20

	Au	CS	Ch	Bl	NonV
SpA	1	1	0	1	
Epi	3	2	0	2	
Vol	0	2	3	0	
NonV					5

> table(SoWrittenA$Rel,SoWrittenA$Soc) Domain Acad, size 19

	Au	CS	Ch	Bl	NonV
SpA	1	0	0	0	
Epi	8	0	0	4	
Vol	0	0	0	1	
NonV					5

1.2. Therefore – domain summaries

> summary(TherefWritten)

Rel	SoC	Modal	Domain
SpA: 22	Au :134	0:168	Fict : 4
Epi :177	CS: 17	1: 27	NonAc :133
Vol : 32	Ch: 23	2: 55	Newsp : 6
NonV : 19	Bl : 57		Acad :107
	(NonV : 19)		

> summary(TherefWrittenF)

Rel	SoC	Modal	Domain Fict, size 4
SpA:0	Au :1	0:3	
Epi :4	CS:0	1:1	
Vol :0	Ch:2	2:0	
NonV :0	Bl :1		
	NonV :0		

> summary(TherefWrittenNA)

Rel	SoC	Modal	Domain NonAc, size 133
SpA:16	Au :71	0:88	
Epi :85	CS:11	1:17	
Vol :18	Ch: 9	2:28	
NonV :14	Bl :28		
	NonV :14		

> summary(TherefWrittenN)

Rel	SoC	Modal	Domain Newsp, size 6
SpA:1	Au :2	0:5	
Epi :4	CS:2	1:1	
Vol :0	Ch:1	2:0	
NonV :1	Bl :0		
	NonV :1		

> summary(TherefWrittenA)

Rel	SoC	Modal	Domain Acad, size 107
SpA: 5	Au :60	0:72	
Epi :84	CS: 4	1: 8	
Vol :14	Ch:11	2:27	
NonV : 4	Bl :28		
	NonV : 4		

1.2.1. *Therefore – Rel versus SoC*

> table(TherefWrittenF$Rel,TherefWrittenF$Soc) Domain Fict, size 4

	Au	CS	Ch	Bl	NonV
SpA	0	0	0	0	
Epi	1	0	2	1	
Vol	0	0	0	0	
NonV					0

> table(TherefWrittenNA$Rel,TherefWrittenNA$Soc) Domain NonAc, size 133

	Au	CS	Ch	Bl	NonV
SpA	13	1	1	1	
Epi	56	7	3	19	
Vol	2	3	5	8	
NonV					14

> table(TherefWrittenN$Rel,TherefWrittenN$Soc) Domain Newsp, size 6

	Au	CS	Ch	Bl	NonV
SpA	0	1	0	0	
Epi	2	1	1	0	
Vol	0	0	0	0	
NonV					1

> table(TherefWrittenA$Rel,TherefWrittenA$Soc) Domain Acad, size 107

	Au	CS	Ch	Bl	NonV
SpA	3	0	0	2	
Epi	55	4	4	21	
Vol	2	0	7	5	
NonV					4

1.3. As a result – domain summaries

> summary(AARWritten) All domains together

Rel	SoC	Modal	Domain sizes
SpA: 0	Au : 27	0:209	Fict : 9
Epi : 52	CS: 7	1: 7	NonAc :157
Vol : 28	Ch: 25	2: 34	Newsp : 29
NonV :170	Bl : 21		Acad : 55
	NonV :170		

> summary(AARWrittenF) Domain Fiction, size 9

Rel	SoC	Modal
SpA:0	Au :0	0:9
Epi :0	CS:0	1:0
Vol :0	Ch:0	2:0
NonV :9	Bl :0	
	NonV :9	

> summary(AARWrittenNA) Domain NonAc, size 157

Rel	SoC	Modal
SpA: 0	Au : 15	0:129
Epi : 31	CS: 6	1: 4
Vol : 21	Ch: 20	2: 24
NonV :105	Bl : 11	
	N :105	

> summary(AARWrittenN) Domain Newsp, size 29

Rel	SoC	Modal
SpA: 0	Au : 4	0:23
Epi : 8	CS: 1	1: 0
Vol : 4	Ch: 3	2: 6
NonV :17	Bl : 4	
	NonV :17	

> summary(AARWrittenA) Domain A, size 55

Rel	SoC	Modal
SpA: 0	Au : 8	0:48
Epi :13	CS: 0	1: 3
Vol : 3	Ch: 2	2: 4
NonV :39	Bl : 6	
	NonV :39	

1.3.1. *As a result - Rel versus SoC*

> table(AARWrittenF$Rel,AARWrittenF$Soc) Domain Fict, size 9

	Au	CS	Ch	Bl	NonV
SpA	0	0	0	0	
Epi	0	0	0	0	
Vol	0	0	0	0	
NonV					9

> table(AARWrittenNA$Rel,AARWrittenNA$Soc) Domain NonAc, size 157

	Au	CS	Ch	Bl	NonV
SpA	0	0	0	0	
Epi	15	5	3	8	
Vol	0	1	17	3	
NonV					105

> table(AARWrittenN$Rel,AARWrittenN$Soc) Domain Newsp, size 29

	Au	CS	Ch	Bl	NonV
SpA	0	0	0	0	
Epi	4	0	0	4	
Vol	0	1	3	0	
NonV					17

> table(AARWrittenA$Rel,AARWrittenA$Soc) Domain Acad, size 55

	Au	CS	Ch	Bl	NonV
SpA	0	0	0	0	
Epi	8	0	0	5	
Vol	0	0	2	1	
NonV					39

1.4. For this reason– domain summaries

> summary(FTRWritten) All domains together

Rel	Soc	Modal	Domain sizes
SpA: 15	Au :104	0:180	Fict : 8
Epi :128	CS: 31	1: 22	NonAc:155
Vol : 85	Ch: 28	2: 48	Newsp : 6
NonV : 22	Bl : 65		Acad : 81
	NonV : 22		

> summary(FTRWrittenF) Domain Fict, size 8

Rel	SoC	Modal
SpA:1	Au :1	0:6
Epi :2	CS:3	1:1
Vol :4	Ch:2	2:1
NonV :1	Bl :1	
	NonV :1	

> summary(FTRWrittenNA) Domain NonAc, size 155

Rel	SoC	Modal
SpA:12	Au :66	0:112
Epi :77	CS:20	1: 15
Vol :52	Ch:18	2: 28
NonV :14	Bl :37	
	NonV :14	

> summary(FTRWrittenN) Domain Newsp, size 6

Rel	SoC	Modal
SpA:0	Au :1	0:5
Epi :4	CS:0	1:0
Vol :2	Ch:0	2:1
NonV :0	Bl :5	
	NonV :0	

> summary(FTRWrittenA) Domain Acad, size 81

Rel	SoC	Modal
SpA: 2	Au :36	0:57
Epi :45	CS: 8	1: 6
Vol :27	Ch: 8	2:18
NonV : 7	Bl :22	
	NonV : 7	

1.4.1. For this reason – Rel versus SoC

> table(FTRWrittenF$Rel,FTRWrittenF$Soc) Domain Fict, size 8

	Au	CS	Ch	Bl	NonV
SpA	0	1	0	0	
Epi	1	0	0	1	
Vol	0	2	2	0	
NonV					1

> table(FTRWrittenNA$Rel,FTRWrittenNA$Soc) Domain NonAc, size 155

	A	CS	Ch	B	NonV
SpA	5	0	1	6	
E	54	9	2	12	
V	7	11	15	19	
NonV					14

> table(FTRWrittenN$Rel,FTRWrittenN$Soc) Domain Newsp, size 6

	A	CS	Ch	B	NonV
SpA	0	0	0	0	
E	1	0	0	3	
V	0	0	0	2	
NonV					0

> table(FTRWrittenA$Rel,FTRWrittenA$Soc) Domain Acad, size 81

	Au	CS	Ch	Bl	NonV
SpA	1	0	0	1	
Epi	32	3	1	9	
Vol	3	5	7	12	
NonV					7

S.II.2. Spoken material

2.1. So – domain summaries

summary(So Spoken) Total = 250

Rel	SoC	Modal	Domain
SpA: 32	Au : 93	0:187	Leis: 114
Epi: 153	CS:107	1: 27	Busn: 51
Vol: 37	Ch: 19	2: 36	Educ: 63
NonV: 28	Bl : 3		Publ: 22
	(NonV : 28)		

So per domain:

summary(SoSpokenL) Leis=Leisure, total = 114

Rel	Soc	Modal
SpA:13	Au :33	0:86
Epi: 57	CS:47	1: 9
Vol: 27	Ch:15	2:19
NonV: 17	B : 2	
	(NonV :17)	

summary(SoSpokenB) Busn=Business, total = 51

Rel	Soc	Modal
SpA:7	Au :20	0:36
Epi: 35	CS:26	1: 7
Vol: 7	Ch: 2	2: 8
NonV: 2	Bl: 1	
	(NonV : 2)	

summary(SoSpokenE) Educ=Education, total = 63

Rel	Soc	Modal
SpA:11	Au : 33	0:52
Epi: 41	CS:19	1: 4
Vol: 2	Ch: 2	2: 7
NonV: 9	B: 0	
	(NonV : 9)	

summary(SoSpokenP) Publ=Public, total = 22

Rel	Soc	Modal
SpA: 1	Au : 7	0:13
Epi:20	CS:15	1: 7
Vol: 1	Ch: 0	2: 2
NonV: 0	Bl: 0	
	(NonV : 0)	

2.1.1. So Spoken – Rel versus SoC

table(SoSpoken$Rel,SoSpoken$Soc) Summed over all domains Total = 250

	Soc				
Rel	Au	CS	Ch	Bl	(NonV)
SpA	23	8	1	0	
Epi	70	76	5	2	
Vol	0	23	13	1	
NonV					28

---

table(SoSpokenL$Rel,SoSpokenL$Soc) Leis=Leisure Total = 114

Rel	Au	CS	Ch	Bl	NonV
SpA	8	4	1	0	
Epi	25	29	2	1	
Vol	0	14	12	1	
NonV					17

table(SoSpokenB$Rel,SoSpokenB$Soc) Busn=Business Total = 51

Rel	Au	CS	Ch	Bl	NonV
SA	6	1	0	0	
Epi	14	18	2	1	
Vol	0	7	0	0	
NonV					2

table(SoSpokenE$Rel,SoSpokenE$Soc) Educ=Education Total = 63

Rel	Au	CS	Ch	Bl	NonV
SpA	9	2	0	0	
Epi	24	16	1	0	
Vol	0	1	1	0	
NonV					9

table(SoSpokenP$Rel,SoSpokenP$Soc) Publ=Public Total = 22

Rel	Au	CS	Ch	Bl	NonV
SpA	0	1	0	0	
Epi	7	13	0	0	
Vol	0	1	0	0	
NonV					0

2.2. Therefore Spoken – domain summaries

> summary(ThereforSpoken) Sample size = 250

Rel	SoC	Modal	Category
SpA: 21	Au :103	0:179	Leis: 36
Epi :181	CS:101	1: 46	Busn: 40
Vol : 36	Ch: 11	2: 25	Educ: 63
NonV : 12	Bl : 23		Publ:111
	(NonV : 12)		

> summary(TherefSpokenL) Subsample Leisure Leis: 36

Rel	Soc	Modal
SpA: 2	Au:12	0:31
Epi :25	CS:14	1: 2
Vol : 7	Ch: 3	2: 3
NonV : 2	Bl : 5	
	NonV : 2	

> summary(TherefSpokenB)) Subsample Business Busn: 40

Rel	Soc	Modal
SpA: 5	Au :12	0:31
Epi :23	CS:19	1: 6
Vol: 8	Ch: 3	2: 3
NonV : 4	Bl : 2	
	NonV : 4	

> summary(TherefSpokenE) Subsample Educational Educ: 63

Rel	Soc	Modal
SpA: 3	Au :32	0:43
Epi :51	CS:20	1:13
Vol : 7	Ch: 3	2: 7
NonV : 2	Bl : 6	
	NonV : 2	

> summary(TherefSpokenP)) Subsample Public Publ: 111

Rel	Soc	Modal
SpA:11	Au :47	0:74
Epi :82	CS:48	1:25
Vol :14	Ch: 2	2:12
NonV : 4	Bl :10	
	NonV : 4	

2.2.1. Therefore Spoken – Rel versus Soc

> table(TherefSpoken$Rel,TherefSpoken$Soc) Total sample, size 250

	Au	CS	Ch	Bl	NonV
SpA	12	8	0	1	
Epi	91	73	2	15	
Vol	0	20	9	7	
NonV					12

> table(TherefSpokenL$Rel,TherefSpokenL$Soc) Subsample Leisure L: 36

	Au	CS	Ch	Bl	NonV
SpA	1	0	0	1	
Epi	11	13	0	1	
Vol	0	1	3	3	
NonV					2

> table(TherefSpokenB$Rel,TherefSpokenB$Soc) Subsample Business, B: 40

	Au	CS	Ch	Bl	NonV
SpA	1	4	0	0	
Epi	11	11	1	0	
Vol	0	4	2	2	
NonV					4

> table(TherefSpokenE$Rel,TherefSpokenE$Soc) Subsample Education, E: 63

	Au	CS	Ch	Bl	NonV
SpA	3	0	0	0	
Epi	29	16	0	6	
Vol	0	4	3	0	
NonV					2

> table(TherefSpokenP$Rel,TherefSpokenP$Soc) Subsample Public, P: 111

	Au	CS	Ch	Bl	NonV
SpA	7	4	0	0	
Epi	40	33	1	8	
Vol	0	11	1	2	
NonV					4

2.3. AAR Spoken – domain summaries

> summary(AARSpoken)

Rel	Soc	Modal	Category
SpA: 3	Au :13	0:59	Leis: 8
Epi:26	CS: 3	1: 5	Busn: 11
Vol:12	Ch: 2	2: 6	Educ: 44
NonV:29	Bl :23		Publ: 7
	NonV :29		

> summary(AARSpokenL) L=Leisure Leis total = 8

Rel	SoC	Modal
SpA:1	Au :2	0:8
Epi :3	CS:0	1:0
Vol :1	Ch:1	2:0
NonV :3	Bl :2	
	(NonV :3)	

> summary(AARSpokenB) B=Business Busn total = 11

Rel	SoC	Modal
SpA:2	Au : 0	0:9
Epi : 2	CS:2	1:1
Vol : 5	Ch:0	2:1
NonV : 2	Bl : 7	
	(NonV :2)	

> summary(AARSpokenE) E=Educational Educ total = 44

Rel	SoC	Modal
SpA: 0	Au :11	0:35
Epi :21	CS: 1	1: 4
Vol : 3	Ch: 1	2: 5
NonV :20	Bl :11	
	NonV :20	

> summary(AARSpokenP) P=Public Publ total = 7

Rel	SoC	Modal
SpA:0	Au :0	0:7
Epi :0	CS:0	1:0
Vol :3	Ch:0	2:0
NonV :4	Bl :3	

2.3.1. AAR Spoken – Rel versus SoC

> table(AARSpoken$Rel,AARSpoken$Soc) All categories Total = 70

Rel \ SoC	Au	CS	Ch	Bl	NonV
SpA	0	0	0	3	
Epi	13	2	1	10	
Vol	0	1	1	10	
NonV					29

> table(AARSpokenL$Rel,AARSpokenL$Soc) L=Leisure L total = 8

Rel \ SoC	Au	CS	Ch	Bl	NonV
SpA	0	0	0	1	
Epi	2	0	0	1	
Vol	0	0	1	0	
NonV					3

> table(AARSpokenB$Rel,AARSpokenB$Soc) B=Business B total = 11

Rel	Au	CS	Ch	Bl	NonV
SpA	0	0	0	2	
Epi	0	1	0	1	
Vol	0	1	0	4	
NonV					2

> table(AARSpokenE$Rel,AARSpokenE$Soc) E=Educational E total = 44

Rel	Au	CS	Ch	Bl	NonV
SpA	0	0	0	0	
Epi	11	1	1	8	
Vol	0	0	0	3	
NonV					20

> table(AARSpokenP$Rel,AARSpokenP$Soc) P=Public P total = 7

Rel	Au	CS	Ch	Bl	NonV
SpA	0	0	0	0	
Epi	0	0	0	0	
Vol	0	0	0	3	
NonV					4

2.4. FTR Spoken – domain summaries

Rel	SoC	Modal	Domain	Total = 6
SpA :0	Au: 0	0:5	Leis:0	
Epi: 1	CS :3	1:0	Busn:2	
Vol: 5	Ch: 2	2:1	Educ:1	
NonV: 0	Bl: 1		Publ:3	

> summary(FTRSpokenL) L total = 0
Empty domain

> summary(FTRSpokenB) Busn total = 2

Rel	SoC	Modal
SpA:0	Au: 0	0:1
Epi: 0	CS :1	1:0
Vol: 2	Ch: 0	2:1
NonV: 0	Bl: 1	

summary(FTRSpokenE) Educ total = 1

Rel	SoC	Modal
SpA:0	Au : 0	0:1
Epi: 1	CS :1	1:0
Vol: 0	Ch:0	2:0
NonV: 0	Bl: 0	

> summary(FTRSpokenP) Publ total = 3

Rel	SoC	Modal
SpA:0	Au :0	0:3
Epi:0	CS :1	1:0
Vol:3	Ch:2	2:0
NonV:0	Bl:0	

2.4.1. FTR Spoken – Rel versus SoC

table(FTRSpoken$Rel,FTRSpoken$Soc) All domains Total = 6

	Au	CS	Ch	Bl	
Epi	0	1	0	0	(Epi, CS) from Domain = Educ
Vol	0	2	2	1	Vol from domains Publ (3) and Busn (2)

SpA and NonV empty

table(FTRSpokenP$Rel,FTRSpokenP$Soc) Domain P=Public Total = 3

	Au	CS	Ch	Bl
Vol	0	1	2	0

SpA, Epi and NonV empty

Section S.III. Predictive analyses. Written data

S.III.1 Relevance and expected sample sizes.

Observed poststratasizes

Conn\Domain	Fict	NonAc	Newsp	Acad	Total
So	75	136	20	19	250
Theref	4	133	6	107	250
AAR	9	157	29	55	250
FTR	8	155	6	81	250

Expected strata sizes, given population strata sizes and target proportions (see below)

Conn\Domain	Fict	NonAc	Newsp	Acad	Total
So	79	135	19	17	250
Theref	7	138	6	99	250
AAR	5	155	21	69	250
FTR	4	149	4	93	250

2. Crude predictive analysis.

Based on the population domain sizes adjusted with respect to relevance proportions as specified on the previous page.

Domain\ % units	So	Theref	AAR	FTR	Tot	Actual domain totals
Fiction	95.8	3.9	0.3	0.1	100%	13 500
NonAc	64.1	31.0	3.5	1.4	100%	34 500
Newsp	82.3	12.9	4.4	0.4	100%	3 800
Acad	24.9	67.7	4.8	2.6	100%	11 300

3. Predictive analysis results, per domain, split by Rel and SoC. Written data.

Non-volitional (NonV) relations are included in the tables below in an NonVxNonV position.
Probability of choice of connective is given as percentage with one decimal digit.
Totally empty combinations of Rel and SoC for the domain in question are shown by a ”–”.

i. Fiction domain, Fict (% units, uncertainty in parentheses)

So

Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	100	99.9 (≥90)	100	100	
Epi	91.3 (≥84)	100 (≥95)	72.6 (≥65)	90.2 (≥82)	
Vol	–	99.9 (≥95)	99.9 (≥95)	–	
NonV					95.8 (93–97)

Therefore (empty, except when Rel = Epistemic)

Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	0.0	0.0	0.0	0.0	
Epi	8.6 (≤16)	0.0 (≤15)	27.4 (≤35)	9.7 (≤18)	
Vol	–	0.0 (≤15)	0.0 (≤5)	–	
NonV					0.0

AAR (empty, except for the NonVol relation type)

Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	0.0	0.0	0.0	0.0	
Epi	0.0	0.0	0.0	0.0	
Vol	–	0.0	0.0	–	
NonV					4.1 (3–7)

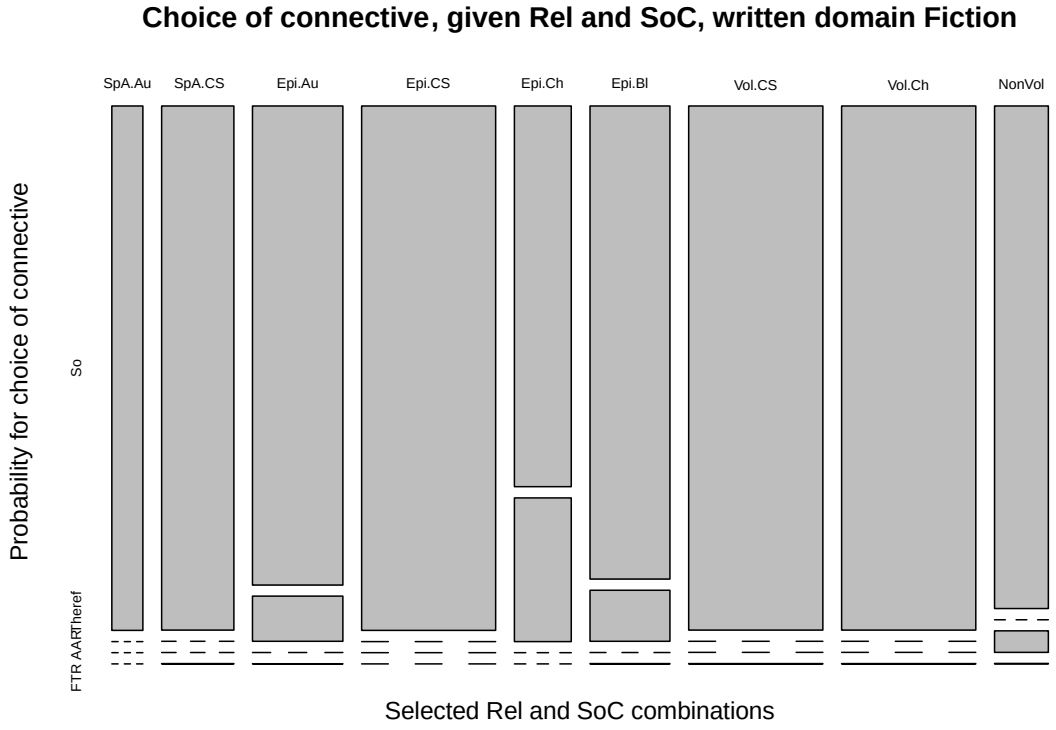
For this reason (FTR)

Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	0.0	0.1	0.0	0.0	
Epi	0.0	0.0	0.0	0.0	
Vol	–	0.1	0.1	–	
NonV					0.1

Sum over the four connectives is 100% for each discourse setting.

Estimated Rel\SoC sizes for domain Fiction:

Rel\SoC	A	CS	Ch	Bl	(NonV)
SpA	516	1206	172	172	
Epi	1508	2238	949	1336	
Vol	0	2239	2239	0	
NonV					898



ii. Non-Academic domain, NonAc

So

Rel\SoC	Au	CS	Ch	Bl	(NonVol)
SpA	71.0 (61–77)	85.8	66.0	93.7	
Epi	44.9 (±5)	79.5 (69–85)	37.5	60.0 (51–66)	
Vol	0.0	90.2 (79–93)	71.6	31.0	
NonVol					67.2

Therefore

Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	28.6 (23–39)	14.2	32.7	5.2	
Epi	51.9 (±5)	18.4 (13–29)	55.7	37.6 (31–47)	
Vol	88.4	8.4 (5–19)	19.7	61.3	
NonVol					18.7

AAR

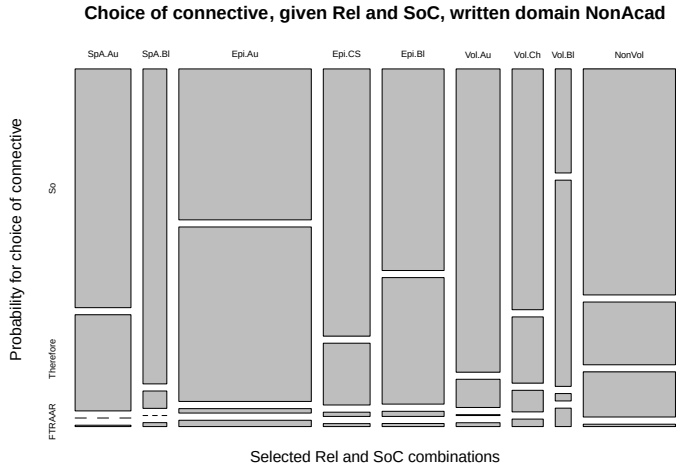
Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	0.0	0.0	0.0	0.0	
Epi	1.3	1.3	5.3	1.5	
V	0.0	0.3	6.4	2.2	
NonVol					13.4

FTR

Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	0.4	0.0	1.2	1.2	
Epi	1.9	0.9	1.4	0.9	
Vol	11.6	1.2	2.2	5.5	
NonVol					0.7

Estimated Rel/SoC-part sizes for domain NonAc

Rel\SoC	Au	CS	Ch	Bl	(NonV)
SpA	3663	568	246	1562	
Epi	8691	3068	433	4067	
Vol	182	2884	2043	1050	
NonVol					6045



iii. Newspaper domain

So

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	100.0	65.6	–	100.0	
Epi	71.3	79.2	0.0	91.3	
Vol	–	98.2	96.4	0.0	
Nonv					81.3

Therefore

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0.0	34.4	–	0.0	
Epi	24.9	20.8	100.0	0.0	
Vol	–	0.0	0.0	0.0	
Nonv					8.5

AAR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0.0	0.0	–	0.0	
Epi	3.5	0.0	0.0	6.7	
Vol	–	1.8	3.6	0.0	
NonV					10.2

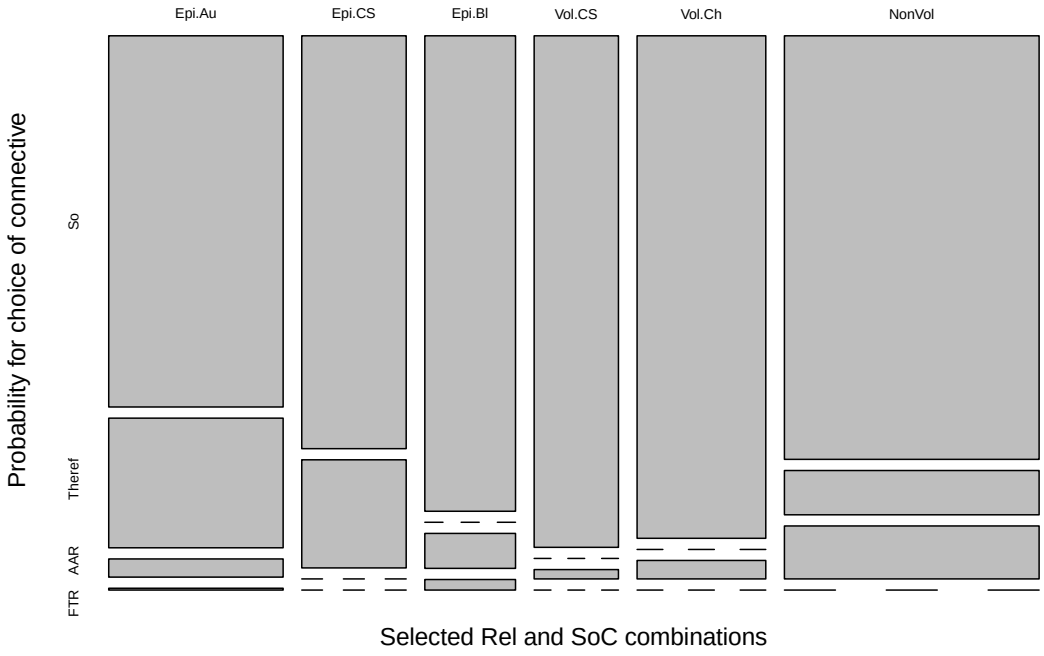
FTR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0.0	0.0	–	0.0	
Epi	0.4	0.0	0.0	2.0	
Vol	–	0.0	0.0	100.0	
NonV					0.0

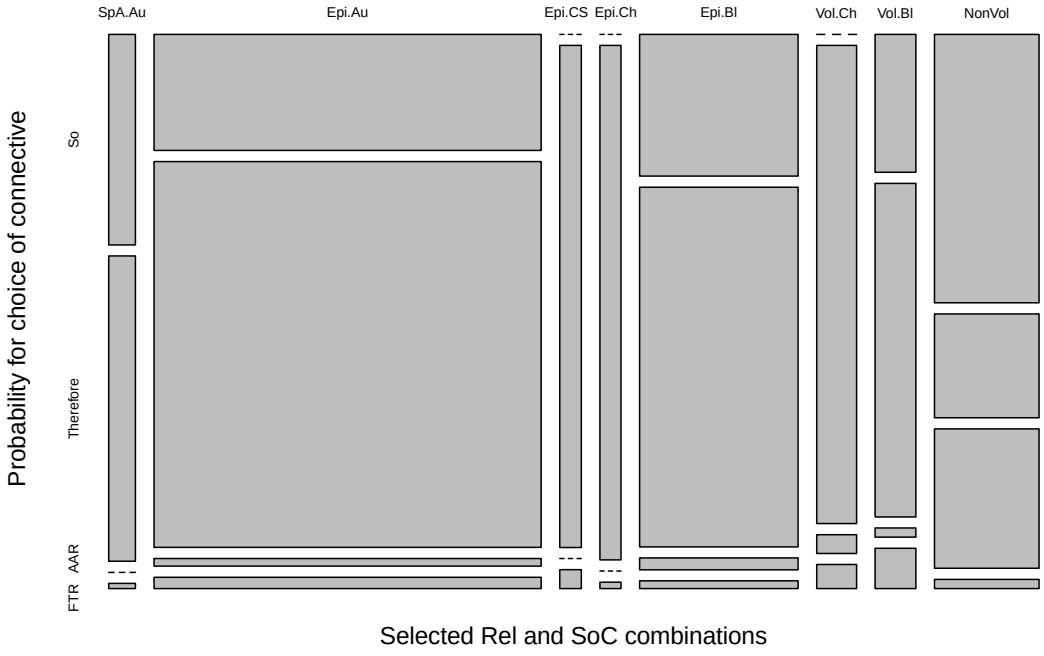
Estimated Rel\SoC sizes for domain Newspaper

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	157	239	0	157	
Epi	660	396	82	344	
Vol	0	319	488	5	
NonV					964

Choice of connective, given Rel and SoC, written domain Newsp



Choice of connective, given Rel and SoC, written domain Academic



Iv. Academic domain Acad (% units, uncertainty in parentheses)

So

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	40.4 (10–68)	–	–	0.0	
Epi	22.3 (±5)	0.0 (≤80)	0.0 (≤80)	27.2 (±11)	
Vol	0.0	0.0	0.0	26.5	
NonV					51.5 (40-60)

Therefore

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	58.6 (31–89)	–	–	97.5	
Epi	74.1 (±5)	96.4 (≥20)	98.8 (≥20)	69.0 (±11)	
Vol	93.0	0.0	91.8	64.0	
NonV					19.9

As a result (AAR)

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0.0	–	–	0.0	
Epi	1.5	0.0	0.0	2.3	
Vol	0.0	0.0	3.6	1.8	
NonV					26.8 (22-34)

For this reason (FTR)

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	1.0	–	–	2.5	
Epi	2.2	3.6	1.2	1.5	
Vol	7.0	100	4.6	7.7	
NonV					1.8 (1-3)

Sum over the four connectives is 100% for each discourse setting

Estimated Rel\SoC sizes for domain Acad

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	367	0	0	147	
Epi	5320	297	290	2179	
Vol	154	18	546	560	
NonV					1438

4. Predictive analysis results, per domain, split by Modality. Written data

Fiction domain, Fict

Connective\Modal	0	1	2
So	96.2	72.3	99.9
Therefore	3.4	27.3	0.0
AAR	0.3	0.0	0.0
FTR	0.1	0.3	0.1

Estimated modal parts of domain Fiction sizes:			
Modality	0	1	2
	11454	476	1551

Aggregated samples sizes:		
0	1	2
82	4	10

Non-Academic domain, NonAc

Connective\Modal	0	1	2
So	68.8	47.4	47.4
Therefore	26.3	49.9	47.0
AAR	3.7	1.1	3.9
FTR	1.3	1.7	1.8

Estimated modal parts of domain NonAcad sizes:			
Modality	0	1	2
	26955	2746	4801

Aggregated samples sizes:		
0	1	2
443	44	94

Newspaper domain, Newsp

Connective\Modal	0	1	2
So	84.3	0.0	80.9
Therefore	11.6	100	0.0
AAR	3.8	0.0	17.9
FTR	0.3	0.0	1.2

Estimated modal parts of domain Newsp sizes:			
Modality	0	1	2
	3534	82	194

Aggregated samples sizes:		
0	1	2
52	1	7

Academic domain, Acad

Connective\Modal	0	1	2
So	23.3	19.2	30.4
Therefore	67.7	74.2	66.1
AAR	6.2	3.8	1.3
FTR	2.7	2.8	2.2

Estimated modal parts of domain Acad sizes:			
Modality	0	1	2
	7616	773	2928

Aggregated samples sizes:		
0	1	2
189	18	55

Section S.IV. Predictive analyses. Spoken data

S.IV.1. Relevance and expected sample sizes.

Observed = actual samples sizes 250, 250, 70 and 6, split on domains

Observed post-strata sizes

Conn\Domain	Leis	Busn	Edu	Publ	Total
So	114	51	63	22	250
Theref	36	40	63	111	250
AAR	8	11	44	7	70
FTR	0	2	1	3	6

Expected strata sizes, given population strata sizes and relevance proportions (see below)

Conn\Domain	Leis	Busn	Edu	Publ	Total
So	56	55	88	50	250
Theref	21	49	61	119	250
AAR	8	11	44	7	70
FTR	0	2	1	3	6

2. Crude inverse analysis, Spoken data.

Population domain sizes for different connectives, Spoken material.

Conn\Domain	Leis	Busn	Edu	Publ	Total	Relevance
So	8244	8102	12896	7256	36498	250/540
Theref	125	285	356	696	1462	250/325
AAR	46	56	122	154	378	70/378
FTR	0	3	3	9	15	6/15

Domain \ % units	So	Theref	AAR	FTR	Tot%	Total per domain
Leis	97	3	0	0	100%	3 900
Busn	94	6	0	0	100%	4 000
Edu	95	4	1	0	100%	6 300
Publ	86	14	0	0	100%	3 900

3. Predictive analysis results, per domain, split by Rel and SoC. Spoken data.

I Public domain

Total sample, summed over all connectives:

Rel/SoC	Au	CS	Ch	Bl	NonV
SpA	7	5	0	0	
Epi	47	46	1	8	
Vol	0	13	3	5	
NonV					8

Zeros correspond to – (missing) in the tables below.

The corresponding subtable for So, only:

Rel/SoC	Au	CS	Ch	Bl	NonV
SpA	0	1	0	0	
Epi	7	13	0	0	
Vol	0	1	0	0	
NonV					0

Total = 22

Analogously for Therefore:

Au	CS	Ch	Bl	NonV
7	4	0	0	
40	33	1	8	
0	11	1	2	
				4

Total: 111

Relevant Publ-stratum size for So, with estimated relevance factor: 7256*(250/540)=3359.3
Relevant P-stratum size for Theref, with estimated relevance factor: 696*(250/325)=535.4
When these Public domain strata are divided according to sample proportions of Rel and SoC combinations, the following results are obtained.

Estimated stratum sizes, split according to sample outcome table, stratum Public:

So

Rel/SoC	Au	CS	Ch	Bl	NonV
SpA	0	153	0	0	
Epi	1069	1985	0	0	
Vol	0	153	0	0	
NonV					0

Total = 3360

Therefore

Au	CS	Ch	Bl	NonV
34	19	0	0	
193	159	5	39	
0	53	5	10	
				19

Total = 536

Due to the large uncertainties in the corresponding estimated probabilities of choice for the Public domain, these probabilities have been rounded below to whole percentage units, also making it easier to see high versus low numbers.

Estimated probabilities for choice of connective, Public domain, in % units:

So

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0	61	–	–	
Epi	52	71	0	0	
Vol	–	36	0	0	
NonV					0

Therefore

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	100	11	–	–	
Epi	15	7	100	100	
Vol	–	26	71	76	
NonV					83

AAR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0	0	–	–	
Epi	0	0	0	0	
Vol	–	0	0	24	
NonV					17

FTR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0	0	–	–	
Epi	0	0	0	0	
Vol	–	1	29	0	
NonV					0

Ii Education domain

Total sample, summed over all connectives:

Rel/SoC	Au	CS	Ch	Bl	NonV
SpA	12	2	0	0	
Epi	64	34	2	14	
Vol	0	5	4	3	
NonV					31

Distribution for choice of connective

So

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	98.5	100	–	–	
Epi	94.3	95.5	99.0	0.0	
Vol	–	84.5	87.9	0.0	
NonV					96.7

Therefore

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	1.5	0.0	–	–	
Epi	5.2	4.4	0.0	76.5	
Vol	–	15.5	12.1	0.0	
NonV					1.0

AAR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA		0.0	0.0	–	–
Epi	0.5	0.1	1.0	23.5	
Vol	–	0.0	0.0	100	
NonV					2.3

FTR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA		0.0	0.0	–	–
Epi	0.0	0.1	0.0	0.0	
Vol	–	0.0	0.0	0.0	
NonV					0.0

Iii Business domain, B

Total sample, summed over all connectives:

Rel/SoC	Au	CS	Ch	Bl	NonV
SpA	7	5	0	2	
Epi	25	30	3	2	
Vol	0	13	2	7	
NonV					8

Zeros correspond to – (missing) in the tables below.

Distribution for choice of connective:

So

Rel\SoC	Au	CS	Ch	Bl	NonV
SA	98.8	77.0	–	0.0	
Epi	95.6	95.6	96.4	98.7	
Vol	–	95.6	0.0	0.0	
NonV					86.0

Therefore

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	1.2	23.0	–	0.0	
Epi	5.5	4.4	3.6	0.0	
Vol	–	4.1	100	68.7	
NonV					12.8

AAR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0.0	0.0	–	100	
Epi	0.0	0.1	0.0	1.3	
Vol	–	0.2	0.0	25.1	
NonV					1.2

FTR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0.0	0.0	–	0.0	
Epi	0.0	0.0	0.0	0.0	
Vol	–	0.2	0.0	6.3	
NonV					0.0

Iv Leisure domain, Leis

Total sample, summed over all connectives:

Rel/SoC	Au	CS	Ch	Bl	NonV
SpA	9	4	1	2	
Epi	38	42	2	3	
Vol	0	15	16	4	
NonV					22

Zeros correspond to – (missing) in the tables below.

Distribution for choice of connective:

So

Rel\SoC	Au	CS	Ch	Bl	NonV
SA	99.0	100	100	0.0	
Epi	96.4	96.5	100	90.1	
Vol	–	99.4	97.8	80.7	
NonV					98.6

Therefore

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	1.0	0.0	0.0	72.8	
Epi	3.4	3.5	0.0	7.2	
Vol	–	0.6	2.0	19.3	
NonV					0.9

AAR

Rel\SoC	Au	CS	Ch	Bl	NonV
SpA	0.0	0.0	0.0	27.2	
Epi	0.2	0.0	0.0	2.7	
Vol	–	0.0	0.2	0.0	
NonV					0.5

(For This Reason) Absent in the Leisure domain

Estimated size of strata parts for the Leisure stratum, all connectives counted and summed:

Rel/SoC	Au	CS	Ch	Bl	NonV
SpA	271	134	33	4	
Epi	868	1006	67	37	
Vol	0	471	411	41	
NonV					577

4. Predictive analysis results, per domain, split by Modality. Spoken data.

Domain Public

Conn\Modal	0	1	2
So	84.4	89.9	84.1
Therefore	15.2	10.1	15.9
AAR	0.3	0.0	0.0
FTR	0.1	0.0	0.0

Domain Educ

Conn\Modal	0	1	2
So	95.7	86.2	94.9
Therefore	3.6	12.9	4.4
AAR	0.7	0.9	0.7
FTR	0.0	0.0	0.1

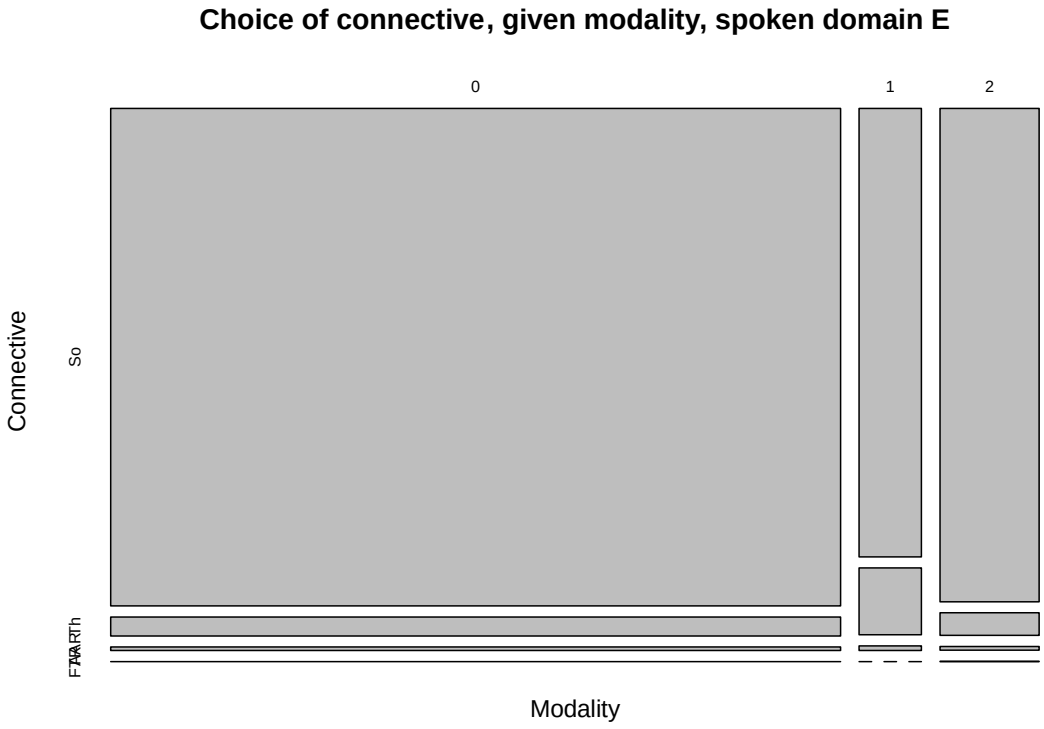
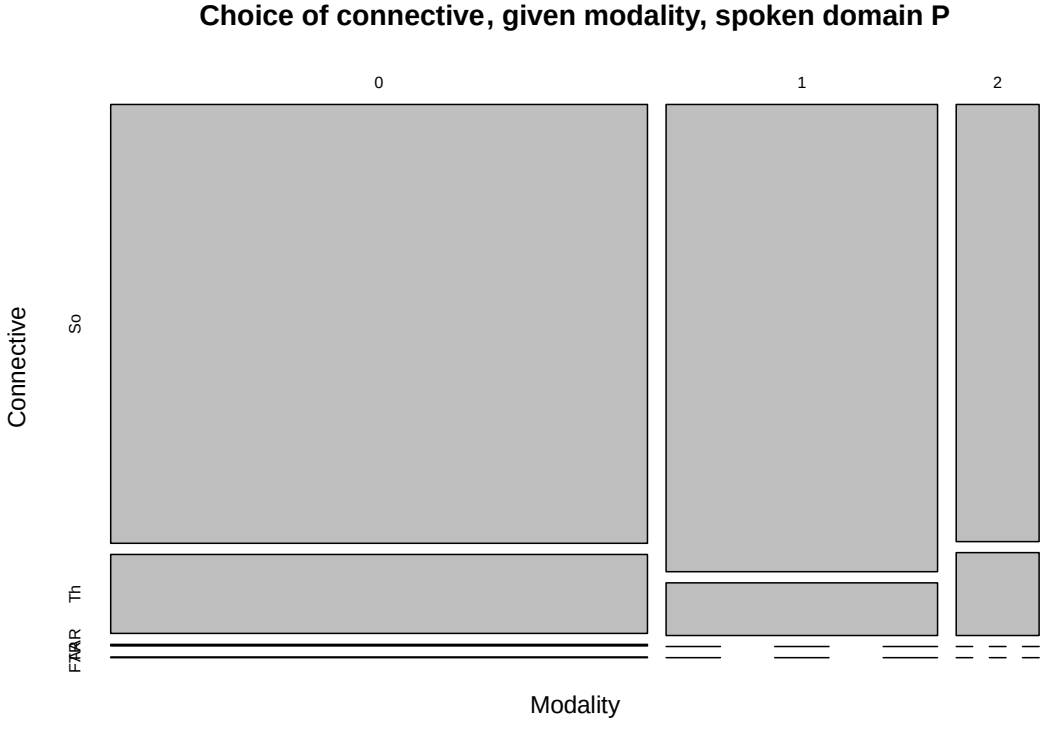
Domain Busn

Conn\Modal	0	1	2
So	93.6	93.8	97.0
Therefore	6.0	6.0	2.7
AAR	0.3	0.2	0.2
FTR	0.0	0.0	0.2

Domain Leis

Conn\Modal	0	1	2
So	96.9	98.3	98.8
Therefore	2.8	1.7	1.2
AAR	0.3	0.0	0.0
FTR	0.0	0.0	0.0

Estimated population stratum sizes, split according to modal type:

Domain	Leis	Busn	Educ	Publ	
Modal type 0	3090	2960	4830	2650	
Modal type 1	270	540	770	870	
Modal type 2	540	500	700	380	
Total	3900	4000	6300	3900	(see section Crude inverse analysis)

Mosaic plots illustrating choice of connective in Education and Public domains, given modality: