Subjectivity of English discourse connectives
SUPPLEMENTARY MATERIAL

List of contents

I. Corpus analysis
II. Corpus sample analysis
    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. Probabilistic analyses. Written data.
    1. Relevance and expected sample sizes.
    2. Crude probabilistic analysis.
    3. Probabilistic analysis results, per domain, split by Rel and SoC.
    4. Probabilistic analysis results, per domain, split by Modality

IV. Probabilistic analyses. Spoken data.
    1. Relevance and expected sample sizes.
    2. Crude probabilistic analysis
    3. Probabilistic analysis results, per domain, split by Rel and SoC
    4. Probabilistic analysis results, per domain, split by Modality

V. Statistical uncertainties, related to RQ1 and RQ2
    1. Uncertainty in domain sizes
    2. Sampling uncertainty, per connective
    3. Uncertainty in the probabilistic approach, whole domains
    4. Uncertainty in the probabilistic approach, domain parts

Section 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 A1 Target instances of the analyzed connectives in the BNC.

Section II. Corpus sample analysis

1. Written material

1.1. So– domain summaries

> summary(SoWritten) All domains together

Rel	Soc	Modal	Domain sizes
SA: 45	A :64	0:209	F : 75
E :106	CS:72	1: 11	NA :136
V : 59	Ch:32	2: 30	N : 20
N : 40	B :42		A : 19
	N :40		

> summary(SoWrittenF) Domain F, size 75

Rel	Soc	Modal
SA:12	A :11	0:64
E :32	CS:33	1: 2
V :26	Ch:18	2: 9
N : 5	B : 8	
	N : 5	

> summary(SoWrittenNA) Domain NA, size 136

Rel	Soc	Modal
SA:29	A :40	0:114
E :55	CS:34	1: 8
V :27	Ch:11	2: 14
N :25	B :26	
	N :25	

> summary(SoWrittenN) Domain N, size 20

Rel	Soc	Modal
SA:3	A :4	0:19
E :7	CS:5	1: 0
V :5	Ch:3	2: 1
N :5	B :3	
	N :5	

> summary(SoWrittenA) Domain A, size 19

Rel	Soc	Modal
SA: 1	A :9	0:12
E :12	CS:0	1: 1
V : 1	Ch:0	2: 6
N : 5	B :5	
	N :5	

1.1.1. So – Rel versus SoC

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

	A	CS	Ch	B	N
SA	3	7	1	1	0
E	8	13	4	7	0
V	0	13	13	0	0
N	0	0	0	0	5

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

	A	CS	Ch	B	N
SA	16	3	1	9	0
E	24	15	1	15	0
V	0	16	9	2	0
N	0	0	0	0	25

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

	A	CS	Ch	B	N
SA	1	1	0	1	0
E	3	2	0	2	0
V	0	2	3	0	0
N	0	0	0	0	5

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

	A	CS	Ch	B	N
SA	1	0	0	0	0
E	8	0	0	4	0
V	0	0	0	1	0
N	0	0	0	0	5

1.2. Therefore – domain summaries

> summary(TherefWritten)

Rel	Soc	Modal	Domain
SA: 22	A :134	0:168	F : 4
E :177	CS: 17	1: 27	NA :133
V : 32	Ch: 23	2: 55	N : 6
N : 19	B : 57		A :107
	N : 19		

> summary(TherefWrittenF)

Rel	Soc	Modal	Domain F, size 4
SA:0	A :1	0:3	
E :4	CS:0	1:1	
V :0	Ch:2	2:0	
N :0	B :1		
	N :0		

> summary(TherefWrittenNA)

Rel	Soc	Modal	Domain NA, size 133
SA:16	A :71	0:88	
E :85	CS:11	1:17	
V :18	Ch: 9	2:28	
N :14	B :28		
	N :14		

> summary(TherefWrittenN)

Rel	Soc	Modal	Domain N, size 6
SA:1	A :2	0:5	
E :4	CS:2	1:1	
V :0	Ch:1	2:0	
N :1	B :0		
	N :1		

> summary(TherefWrittenA)

Rel	Soc	Modal	Domain A, size 107
SA: 5	A :60	0:72	
E :84	CS: 4	1: 8	
V :14	Ch:11	2:27	
N : 4	B :28		
	N : 4		

1.2.1. *Therefore* – *Rel versus SoC*

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

	A	CS	Ch	B	N
SA	0	0	0	0	0
E	1	0	2	1	0
V	0	0	0	0	0
N	0	0	0	0	0

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

	A	CS	Ch	B	N
SA	13	1	1	1	0
E	56	7	3	19	0
V	2	3	5	8	0
N	0	0	0	0	14

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

	A	CS	Ch	B	N
SA	0	1	0	0	0
E	2	1	1	0	0
V	0	0	0	0	0
N	0	0	0	0	1

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

	A	CS	Ch	B	N
SA	3	0	0	2	0
E	55	4	4	21	0
V	2	0	7	5	0
N	0	0	0	0	4

1.3. As a result – domain summaries

> summary(AARWritten) All domains together

Rel	Soc	Modal	Domain sizes
SA: 0	A : 27	0:209	F : 9
E : 52	CS: 7	1: 7	NA :157
V : 28	Ch: 25	2: 34	N : 29
N :170	B : 21		A : 55
	N :170		

> summary(AARWrittenF) Domain F, size 9

Rel	Soc	Modal
SA:0	A :0	0:9
E :0	CS:0	1:0
V :0	Ch:0	2:0
N :9	B :0	
	N :9	

> summary(AARWrittenNA) Domain NA, size 157

Rel	Soc	Modal
SA: 0	A : 15	0:129
E : 31	CS: 6	1: 4
V : 21	Ch: 20	2: 24
N :105	B : 11	
	N :105	

> summary(AARWrittenN) Domain N, size 29

Rel	Soc	Modal
SA: 0	A : 4	0:23
E : 8	CS: 1	1: 0
V : 4	Ch: 3	2: 6
N :17	B : 4	
	N :17	

> summary(AARWrittenA) Domain A, size 55

Rel	Soc	Modal
SA: 0	A : 8	0:48
E :13	CS: 0	1: 3
V : 3	Ch: 2	2: 4
N :39	B : 6	
	N :39	

1.3.1. *As a result - Rel versus SoC*

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

	A	CS	Ch	B	N
SA	0	0	0	0	0
E	0	0	0	0	0
V	0	0	0	0	0
N	0	0	0	0	9

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

	A	CS	Ch	B	N
SA	0	0	0	0	0
E	15	5	3	8	0
V	0	1	17	3	0
N	0	0	0	0	105

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

	A	CS	Ch	B	N
SA	0	0	0	0	0
E	4	0	0	4	0
V	0	1	3	0	0
N	0	0	0	0	17

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

	A	CS	Ch	B	N
SA	0	0	0	0	0
E	8	0	0	5	0
V	0	0	2	1	0
N	0	0	0	0	39

1.4. For this reason– domain summaries

```
> summary(FTRWritten)        All domains together
 Rel          Soc          Modal        Domain sizes
SA: 15        A :104       0:180        F  :  8
 E :128       CS: 31       1: 22        NA :155
 V : 85       Ch: 28       2: 48        N  :  6
 N : 22       B : 65                    A  : 81
              N : 22
```

```
> summary(FTRWrittenF)              Domain F,   size 8
 Rel          Soc          Modal
 SA:1         A :1         0:6
 E :2         CS:3         1:1
 V :4         Ch:2         2:1
 N :1         B :1
              N :1
```

```
> summary(FTRWrittenNA)          Domain NA,   size 155
 Rel          Soc          Modal
 SA:12        A :66        0:112
 E :77        CS:20        1: 15
 V :52        Ch:18        2: 28
 N :14        B :37
              N :14
```

```
> summary(FTRWrittenN)          Domain N,  size 6
 Rel          Soc          Modal
 SA:0         A :1         0:5
 E :4         CS:0         1:0
 V :2         Ch:0         2:1
 N :0         B :5
              N :0
```

```
> summary(FTRWrittenA)          Domain A,  size 81
 Rel          Soc          Modal
 SA: 2        A :36        0:57
 E :45        CS: 8        1: 6
 V :27        Ch: 8        2:18
 N : 7        B :22
              N : 7
```

1.4.1. For this reason – Rel versus SoC

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

	A	CS	Ch	B	N
SA	0	1	0	0	0
E	1	0	0	1	0
V	0	2	2	0	0
N	0	0	0	0	1

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

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

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

	A	CS	Ch	B	N
SA	0	0	0	0	0
E	1	0	0	3	0
V	0	0	0	2	0
N	0	0	0	0	0

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

	A	CS	Ch	B	N
SA	1	0	0	1	0
E	32	3	1	9	0
V	3	5	7	12	0
N	0	0	0	0	7

2. Spoken material

2.1. So – domain summaries

summary(SoSpoken) Total = 250

Rel	Soc	Modal	Domain
SA:32	A : 93	0:187	L:114
E: 153	CS:107	1: 27	B: 51
V: 37	Ch: 19	2: 36	E: 63
N: 28	B : 3		P: 22
	N : 28		

So per domain:

summary(SoSpokenL) L=Leisure, total = 114

Rel	Soc	Modal
SA:13	A :33	0:86
E: 57	CS:47	1: 9
V: 27	Ch:15	2:19
N: 17	B : 2	
	N :17	

summary(SoSpokenB) B=Business, total = 51

Rel	Soc	Modal
SA:7	A :20	0:36
E: 35	CS:26	1: 7
V: 7	Ch: 2	2: 8
N: 2	B: 1	
	N : 2	

summary(SoSpokenE) E=Education, total = 63

Rel	Soc	Modal
SA:11	A : 33	0:52
E: 41	CS:19	1: 4
V: 2	Ch: 2	2: 7
N: 9	B: 0	
	N : 9	

summary(SoSpokenP) P=Public, total = 22

Rel	Soc	Modal
SA: 1	A : 7	0:13
E:20	CS:15	1: 7
V: 1	Ch: 0	2: 2
N: 0	B: 0	
	N : 0	

2.1.1. So Spoken – Rel versus SoC

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

Rel \ Soc	A	CS	Ch	B	N
SA	23	8	1	0	0
E	70	76	5	2	0
V	0	23	13	1	0
N	0	0	0	0	28

---

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

Rel	A	CS	Ch	B	N
SA	8	4	1	0	0
E	25	29	2	1	0
V	0	14	12	1	0
N	0	0	0	0	17

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

Rel	A	CS	Ch	B	N
SA	6	1	0	0	0
E	14	18	2	1	0
V	0	7	0	0	0
N	0	0	0	0	2

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

Rel	A	CS	Ch	B	N
SA	9	2	0	0	0
E	24	16	1	0	0
V	0	1	1	0	0
N	0	0	0	0	9

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

Rel	A	CS	Ch	B	N
SA	0	1	0	0	0
E	7	13	0	0	0
V	0	1	0	0	0
N	0	0	0	0	0

2.2. Therefore Spoken – domain summaries

> summary(ThereforSpoken) Sample size = 250

Rel	Soc	Modal	Category
SA: 21	A :103	0:179	L: 36
E :181	CS:101	1: 46	B: 40
V : 36	Ch: 11	2: 25	E: 63
N : 12	B : 23		P:111
	N : 12		

> summary(TherefSpokenL) Subsample Leisure L: 36

Rel	Soc	Modal
SA: 2	A :12	0:31
E :25	CS:14	1: 2
V : 7	Ch: 3	2: 3
N : 2	B : 5	
	N : 2	

> summary(TherefSpokenB)) Subsample Business B: 40

Rel	Soc	Modal
SA: 5	A :12	0:31
E :23	CS:19	1: 6
V : 8	Ch: 3	2: 3
N : 4	B : 2	
	N : 4	

> summary(TherefSpokenE) Subsample Educational E: 63

Rel	Soc	Modal
SA: 3	A :32	0:43
E :51	CS:20	1:13
V : 7	Ch: 3	2: 7
N : 2	B : 6	
	N : 2	

> summary(TherefSpokenP)) Subsample Public P: 111

Rel	Soc	Modal
SA:11	A :47	0:74
E :82	CS:48	1:25
V :14	Ch: 2	2:12
N : 4	B :10	
	N : 4	

2.2.1. Therefore Spoken – Rel versus Soc

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

	A	CS	Ch	B	N
SA	12	8	0	1	0
E	91	73	2	15	0
V	0	20	9	7	0
N	0	0	0	0	12

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

	A	CS	Ch	B	N
SA	1	0	0	1	0
E	11	13	0	1	0
V	0	1	3	3	0
N	0	0	0	0	2

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

	A	CS	Ch	B	N
SA	1	4	0	0	0
E	11	11	1	0	0
V	0	4	2	2	0
N	0	0	0	0	4

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

	A	CS	Ch	B	N
SA	3	0	0	0	0
E	29	16	0	6	0
V	0	4	3	0	0
N	0	0	0	0	2

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

	A	CS	Ch	B	N
SA	7	4	0	0	0
E	40	33	1	8	0
V	0	11	1	2	0
N	0	0	0	0	4

2.3. AAR Spoken – domain summaries

> summary(AARSpoken)

Rel	Soc	Modal	Category
SA: 3	A :13	0:59	L: 8
E:26	CS: 3	1: 5	B: 11
V:12	Ch: 2	2: 6	E: 44
N:29	B :23		P: 7
	N :29		

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

Rel	Soc	Modal
SA:1	A :2	0:8
E :3	CS:0	1:0
V :1	Ch:1	2:0
N :3	B :2	
	N :3	

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

Rel	Soc	Modal
SA:2	A : 0	0:9
E : 2	CS:2	1:1
V : 5	Ch:0	2:1
N : 2	B : 7	
	N :2	

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

Rel	Soc	Modal
SA: 0	A :11	0:35
E :21	CS: 1	1: 4
V : 3	Ch: 1	2: 5
N :20	B :11	
	N :20	

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

Rel	Soc	Modal
SA:0	A :0	0:7
E :0	CS:0	1:0
V :3	Ch:0	2:0
N :4	B :3	
	N :4	

2.3.1. AAR Spoken – Rel versus SoC

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

SoC					
Rel	A	CS	Ch	B	N
SA	0	0	0	3	0
E	13	2	1	10	0
V	0	1	1	10	0
N	0	0	0	0	29

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

SoC					
Rel	A	CS	Ch	B	N
SA	0	0	0	1	0
E	2	0	0	1	0
V	0	0	1	0	0
N	0	0	0	0	3

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

Rel	A	CS	Ch	B	N
SA	0	0	0	2	0
E	0	1	0	1	0
V	0	1	0	4	0
N	0	0	0	0	2

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

Rel	A	CS	Ch	B	N
SA	0	0	0	0	0
E	11	1	1	8	0
V	0	0	0	3	0
N	0	0	0	0	20

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

Rel	A	CS	Ch	B	N
SA	0	0	0	0	0
E	0	0	0	0	0
V	0	0	0	3	0
N	0	0	0	0	4

2.4. FTR Spoken – domain summaries

Rel	Soc	Modal	Domain	Total = 6
SA:0	A : 0	0:5	L:0	
E: 1	CS :3	1:0	B:2	
V: 5	Ch:2	2:1	E:1	
N: 0	B: 1		P:3	
	N : 0			

> summary(FTRSpokenL) L total = 0

Empty domain

> summary(FTRSpokenB) B total = 2

Rel	Soc	Modal
SA:0	A :0	0:1
E: 0	CS :1	1:0
V: 2	Ch:0	2:1
N: 0	B:1	
	N :0	

summary(FTRSpokenE) E total = 1

Rel	Soc	Modal
SA:0	A : 0	0:1
E: 1	CS :1	1:0
V: 0	Ch:0	2:0
N: 0	B: 0	
	N :0	

> summary(FTRSpokenP) P total = 3

Rel	Soc	Modal
SA:0	A :0	0:3
E:0	CS :1	1:0
V:3	Ch:2	2:0
N:0	B:0	
	N :0	

2.4.1. FTR Spoken – Rel versus SoC

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

	A	CS	Ch	B	N	
E	0	1	0	0	0	(E, CS) from Domain = E
V	0	2	2	1	0	V from domains P (3) and B (2)

SA and N empty

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

	A	CS	Ch	B	N
V	0	1	2	0	0

SA, E and N empty

Section III. Probabilistic analyses. Written data

1. Relevance and expected sample sizes.

Observed poststratasizes

Conn\Domain	Fict	NonA	News	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	NonA	News	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 probabilistic analysis.

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

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

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

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, F (% units, uncertainty in parentheses)

So

Rel\SoC	A	CS	Ch	B	N
SA	100	99.9 (≥90)	100	100	–
E	91.3 (≥84)	100 (≥95)	72.6 (≥65)	90.2 (≥82)	–
V	–	99.9 (≥95)	99.9 (≥95)	–	–
N	–	–	–	–	95.8 (93–97)

Therefore (empty, except when Rel = Epistemic)

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	0.0	0.0	–
E	8.6 (≤16)	0.0 (≤15)	27.4 (≤35)	9.7 (≤18)	–
V	–	0.0 (≤15)	0.0 (≤5)	–	–
N	–	–	–	–	0.0

AAR (empty, except for the (N,N) combination)

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	0.0	0.0	–
E	0.0	0.0	0.0	0.0	–
V	–	0.0	0.0	–	–
N	–	–	–	–	4.1 (3–7)

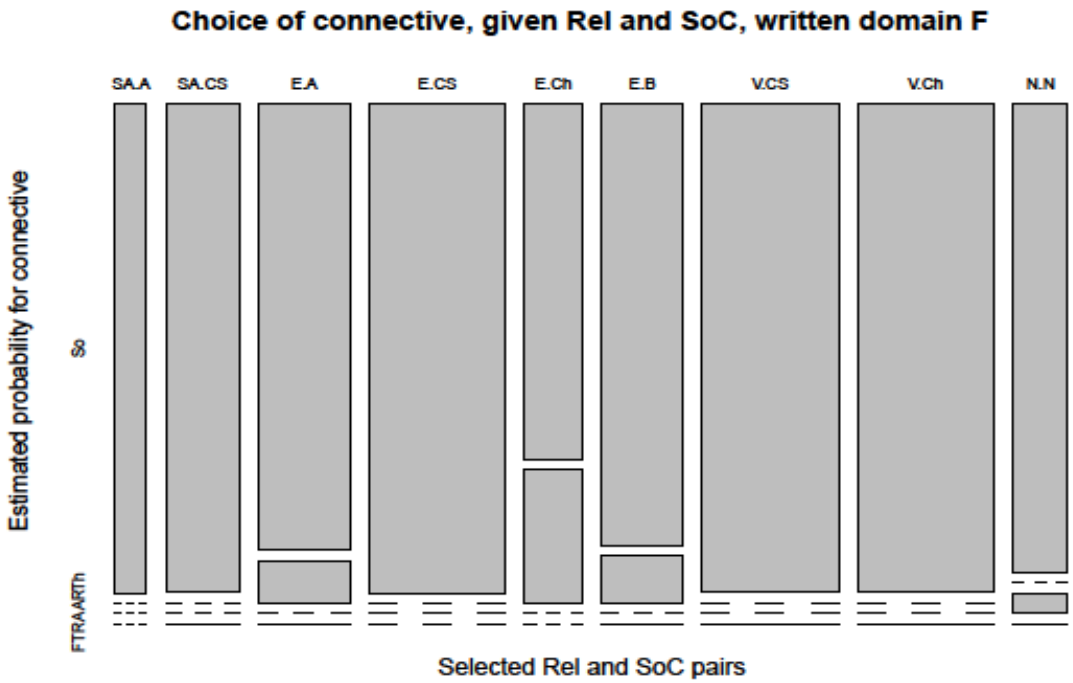
FTR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.1	0.0	0.0	–
E	0.0	0.0	0.0	0.0	–
V	–	0.1	0.1	–	–
N	–	–	–	–	0.1

Sum over connectives is 100% for each discourse setting

Estimated Rel/SoC-part sizes for domain F

Rel\SoC	A	CS	Ch	B	N
SA	516	1206	172	172	0
E	1508	2238	949	1336	0
V	0	2239	2239	0	0
N	0	0	0	0	898



ii. Non-Academic domain, NA

So

Rel\SoC	A	CS	Ch	B	N
SA	71.0	85.8	66.0	93.7	–
E	44.9	79.5	37.5	60.0	–
V	0.0	90.2	71.6	31.0	–
N	–	–	–	–	67.2

Therefore

Rel\SoC	A	CS	Ch	B	N
SA	28.6	14.2	32.7	5.2	–
E	51.9	18.4	55.7	37.6	–
V	88.4	8.4	19.7	61.3	–
N	–	–	–	–	18.7

AAR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	0.0	0.0	–
E	1.3	1.3	5.3	1.5	–
V	0.0	0.3	6.4	2.2	–
N	–	–	–	–	13.4

FTR

Rel\SoC	A	CS	Ch	B	N
SA	0.4	0.0	1.2	1.2	–
E	1.9	0.9	1.4	0.9	–
V	11.6	1.2	2.2	5.5	–
N	–	–	–	–	0.7

Estimated Rel/SoC-part sizes for domain NA

Rel\SoC	A	CS	Ch	B	N
SA	3663	568	246	1562	0
E	8691	3068	433	4067	0
V	182	2884	2043	1050	0
N	0	0	0	0	6045

iii. Newspaper domain

So

Rel\SoC	A	CS	Ch	B	N
SA	100.0	65.6	–	100.0	–
E	71.3	79.2	0.0	91.3	–
V	–	98.2	96.4	0.0	–
N	–	–	–	–	81.3

Therefore

Rel\SoC	A	CS	Ch	B	N
SA	0.0	34.4	–	0.0	–
E	24.9	20.8	100.0	0.0	–
V	–	0.0	0.0	0.0	–
N	–	–	–	–	8.5

AAR

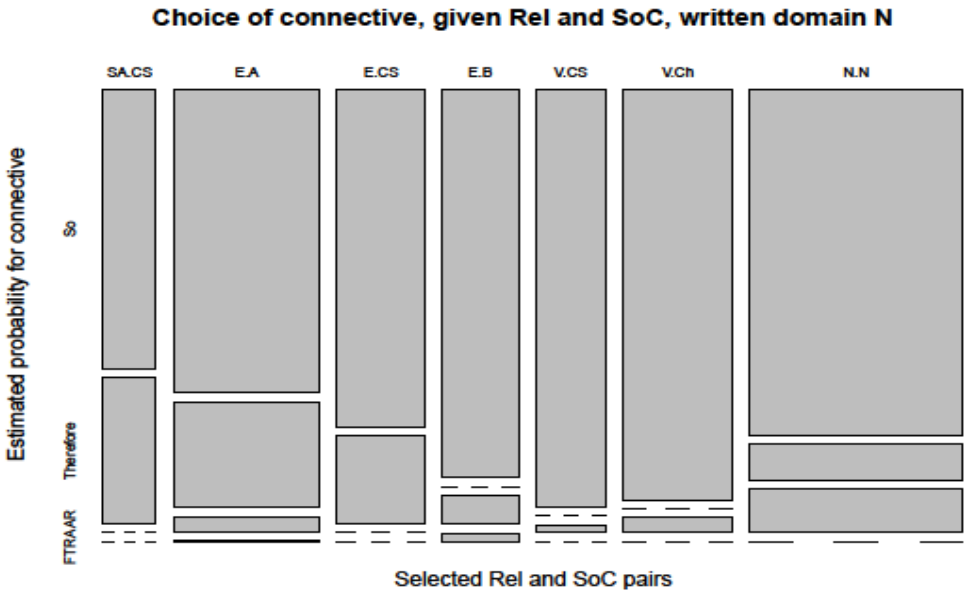
Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	–	0.0	–
E	3.5	0.0	0.0	6.7	–
V	–	1.8	3.6	0.0	–
N	–	–	–	–	10.2

FTR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	–	0.0	–
E	0.4	0.0	0.0	2.0	–
V	–	0.0	0.0	100.0	–
N	–	–	–	–	0.0

Estimated Rel/SoC-part sizes for domain N

Rel\SoC	A	CS	Ch	B	N
SA	157	239	0	157	0
E	660	396	82	344	0
V	0	319	488	5	0
N	0	0	0	0	964


Choice of connective, given Rel and SoC, written domain N
SA.CS
E.A
E.CS
E.B
V.CS
V.Ch
N.N
Estimated probability for connective
So
Therefore
FTRAAR
Selected Rel and SoC pairs

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

So

Rel\SoC	A	CS	Ch	B	N
SA	40.4 (10–68)	–	–	0.0	–
E	22.3 (±5)	0.0 (≤80)	0.0 (≤80)	27.2 (±11)	–
V	0.0	0.0	0.0	26.5	–
N	–	–	–	–	51.5 (40-60)

Therefore

Rel\SoC	A	CS	Ch	B	N
SA	58.6 (31–89)	–	–	97.5	–
E	74.1 (±5)	96.4 (≥20)	98.8 (≥20)	69.0 (±11)	–
V	93.0	0.0	91.8	64.0	–
N	–	–	–	–	19.9

AAR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	–	–	0.0	–
E	1.5	0.0	0.0	2.3	–
V	0.0	0.0	3.6	1.8	–
N	–	–	–	–	26.8 (22-34)

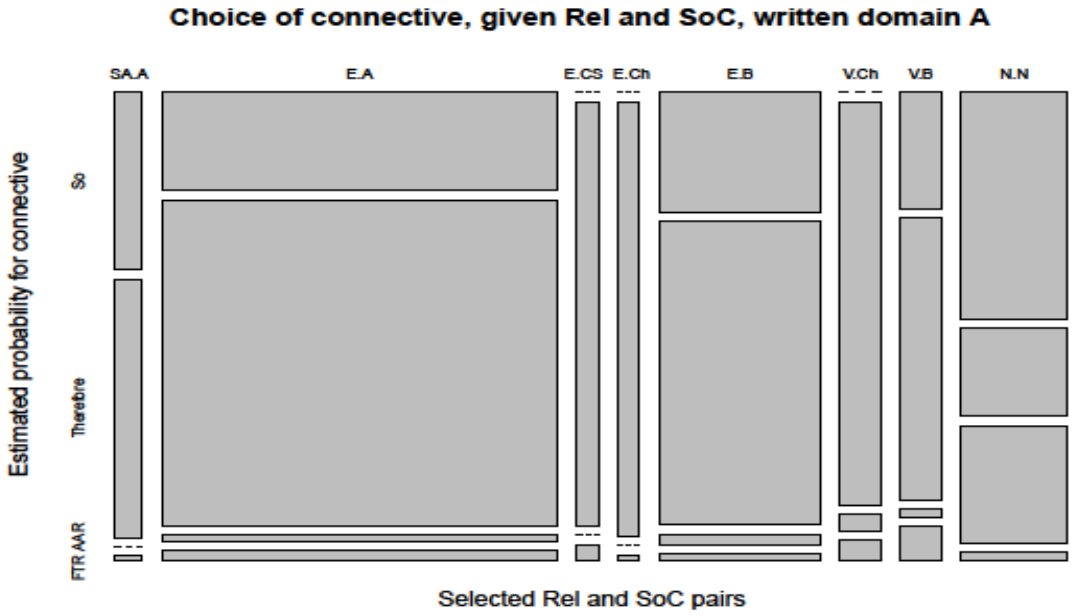
FTR

Rel\SoC	A	CS	Ch	B	N
SA	1.0	–	–	2.5	–
E	2.2	3.6	1.2	1.5	–
V	7.0	100	4.6	7.7	–
N	–	–	–	–	1.8 (1-3)

Sum over connectives is 100% for each discourse setting

Estimated Rel/SoC-part sizes for domain A

Rel\SoC	A	CS	Ch	B	N
SA	367	0	0	147	0
E	5320	297	290	2179	0
V	154	18	546	560	0
N	0	0	0	0	1438


Choice of connective, given Rel and SoC, written domain A
SA.A
E.A
E.CS
E.Ch
E.B
V.Ch
V.B
N.N
So
Therefore
FTR.AAR
Estimated probability for connective
Selected Rel and SoC pairs

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

Fiction domain, F

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 F sizes:			
Modality	0	1	2
	11454	476	1551

Aggregated samples sizes:		
0	1	2
82	4	10

Non-Academic domain, NA

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 NA sizes:			
Modality	0	1	2
	26955	2746	4801

Aggregated samples sizes:		
0	1	2
443	44	94

Newspaper domain, N

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 N sizes:			
Modality	0	1	2
	3534	82	194

Aggregated samples sizes:		
0	1	2
52	1	7

Academic domain, A

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 A sizes:			
Modality	0	1	2
	7616	773	2928

Aggregated samples sizes:		
0	1	2
189	18	55

Section IV. Probabilistic analyses. Spoken data

1. Relevance and expected sample sizes.

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

Observed poststratasizes

Conn\Domain	Leis	Bus	Ed	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	Bus	Ed	Publ	Total
So	56	55	88	50	250 (249)
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	Bus	Ed	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	or /387?
FTR	0	3	3	9	15		6/15	

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

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

i. Public domain

Total sample, summed over all connectives:

Rel/SoC	A	CS	Ch	B	N
SA	7	5	0	0	0
E	47	46	1	8	0
V	0	13	3	5	0
N	0	0	0	0	8

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

The corresponding subtable for So, only:

Rel/SoC	A	CS	Ch	B	N
SA	0	1	0	0	0
E	7	13	0	0	0
V	0	1	0	0	0
N	0	0	0	0	0

Total = 22

Analogously for Therefore:

A	CS	Ch	B	N
7	4	0	0	0
40	33	1	8	0
0	11	1	2	0
0	0	0	0	4

Total: 111

Relevant P-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 P-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 P:

<u>So</u>

Rel/SoC	A	CS	Ch	B	N
SA	0	153	0	0	0
E	1069	1985	0	0	0
V	0	153	0	0	0
N	0	0	0	0	0

Total = 3360 (difference due to roundings)

<u>Therefore</u>

A	CS	Ch	B	N
34	19	0	0	0
193	159	5	39	0
0	53	5	10	0
0	0	0	0	19

Analogously: 536

Due to the large uncertainties in the corresponding estimated probabilities for the P domain, they have been rounded 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	A	CS	Ch	B	N
SA	0	61	–	–	–
E	52	71	0	0	–
V	–	36	0	0	–
N	–	–	–	–	0

Therefore

Rel\SoC	A	CS	Ch	B	N
SA	100	11	–	–	–
E	15	7	100	100	–
V	–	26	71	76	–
N	–	–	–	–	83

AAR

Rel\SoC	A	CS	Ch	B	N
SA	0	0	–	–	–
E	0	0	0	0	–
V	–	0	0	24	–
N	–	–	–	–	17

FTR

Rel\SoC	A	CS	Ch	B	N
SA	0	0	–	–	–
E	0	0	0	0	–
V	–	1	29	0	–
N	–	–	–	–	0

ii. Education domain

Total sample, summed over all connectives:

Rel/SoC	A	CS	Ch	B	N
SA	12	2	0	0	0
E	64	34	2	14	0
V	0	5	4	3	0
N	0	0	0	0	31

Distribution for choice of connective

So

Rel\SoC	A	CS	Ch	B	N
SA	98.5	100	–	–	
E	94.3	95.5	99.0	0.0	
V	–	84.5	87.9	0.0	
N					96.7

Therefore

Rel\SoC	A	CS	Ch	B	N
SA	1.5	0.0	–	–	
E	5.2	4.4	0.0	76.5	
V	–	15.5	12.1	0.0	
N					1.0

AAR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	–	–	
E	0.5	0.1	1.0	23.5	
V	–	0.0	0.0	100	
N					2.3

FTR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	–	–	
E	0.0	0.1	0.0	0.0	
V	–	0.0	0.0	0.0	
N					0.0

iii. Business domain, B

Total sample, summed over all connectives:

Rel/SoC	A	CS	Ch	B	N
SA	7	5	0	2	0
E	25	30	3	2	0
V	0	13	2	7	0
N	0	0	0	0	8

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

Distribution for choice of connective

So

Rel\SoC	A	CS	Ch	B	N
SA	98.8	77.0	–	0.0	
E	95.6	95.6	96.4	98.7	
V	–	95.6	0.0	0.0	
N					86.0

Therefore

Rel\SoC	A	CS	Ch	B	N
SA	1.2	23.0	–	0.0	
E	5.5	4.4	3.6	0.0	
V	–	4.1	100	68.7	
N					12.8

AAR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	–	100	
E	0.0	0.1	0.0	1.3	
V	–	0.2	0.0	25.1	
N					1.2

FTR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	–	0.0	
E	0.0	0.0	0.0	0.0	
V	–	0.2	0.0	6.3	
N					0.0

iv. Leisure domain, L

Total sample, summed over all connectives:

Rel/SoC	A	CS	Ch	B	N
SA	9	4	1	2	0
E	38	42	2	3	0
V	0	15	16	4	0
N	0	0	0	0	22

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

Distribution for choice of connective:

So

Rel\SoC	A	CS	Ch	B	N
SA	99.0	100	100	0.0	
E	96.4	96.5	100	90.1	
V	–	99.4	97.8	80.7	
N					98.6

Therefore

Rel\SoC	A	CS	Ch	B	N
SA	1.0	0.0	0.0	72.8	
E	3.4	3.5	0.0	7.2	
V	–	0.6	2.0	19.3	
N					0.9

AAR

Rel\SoC	A	CS	Ch	B	N
SA	0.0	0.0	0.0	27.2	
E	0.2	0.0	0.0	2.7	
V	–	0.0	0.2	0.0	
N					0.5

FTR
Absent in the L domain

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

Rel/SoC	A	CS	Ch	B	N
SA	271	134	33	4	0
E	868	1006	67	37	0
V	0	471	411	41	0
N	0	0	0	0	577

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

Domain P

Connective\Modality	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 E

Connective\Modality	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 B

Connective\Modality	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 L

Connective\Modality	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

Supplementary material, Section V. Statistical uncertainties related to RQ1 and RQ2

V.1. Uncertainty in domain sizes

For each connective, 'crude' domain sizes, counting all instances, are known. Domain sizes for target instances are not known, however, but must be estimated. This is done by reducing the crude number by the estimated target proportion for the connective in question (and sometimes for the domain in question). The data for the estimated target proportions are given in Section I. They are more or less uncertain. Some adjustments have been made when they have appeared to differ between domains (in particular for connective *so*, see below). In the probabilistic approach (inverse analyses, RQ2), the target proportions play an important role, since they are different for different connectives, whereas they are not at all important for the sample studies undertaken per connective (RQ1). Here are two examples:

(1) *So* Spoken, 250 target instances of 540; $250/540 = 0.46$, or 46%;

(2) *Therefore* Written, 250 target instances of 276; $250/276 = 0.91$, or 91%.

The uncertainty in these percentages, from only sampling randomness, when expressed as standard errors, *s.e.*, is in both cases 2 percentage units, and we write 0.46 ± 0.02 and 0.91 ± 0.02. The corresponding relative errors in the estimated target proportions are $\pm 4\%$ and $\pm 2\%$ of the proportions, respectively. This is not much, and can often be neglected in comparison with other sources of error and uncertainty.

We can often think of the standard error as a measure of confidence, such that with about 95% confidence the true value is in the interval 'estimate $\pm 2 \times$ *s.e.*'. This comes from the normal distribution. The corresponding interval with one *s.e.* has then about 70% confidence.

For *so* Written, it was realized that the target proportion no doubt varied between domains. This is probably not exclusively the case for *so* Written. Such effects generate an additional error. Substantial such effects for a connective will be observed when comparing what we should expect, if the target proportion did not vary, with the actually observed domain strata sizes from the sampling, but it is not possible to compensate for them with precision, due to the sampling uncertainty. The likely presence of some such systematic errors should be kept in mind when judging the results of the inverse analyses. Here is an example:

(3) For *for this reason* (FTR), Written, we expect that (only) 4 items of the sample of 250 should fall in the Fiction domain, as judged from the crude domain sizes and assuming the same target proportion in all domains. The actual sample size for the Fiction domain was 8, and the probability of getting such a large sample size as 8 (or larger), when 4 was expected, is only about $0.05 = 5\%$. This small probability is an indication that there might be reason to doubt the assumption of domain-independent target proportion for FTR in the population. However, since FTR is generally quite an infrequent connective, and particularly so in Fiction, it was judged unnecessary to revise the assumption for FTR. On the other hand, it was revised for *so*.

V.2. Sampling uncertainty, per connective

This uncertainty directly affects the answers to questions of type RQ1, but also indirectly RQ2 analyses (see later).

Sample sizes of 250 are not small, but when we found it necessary to partition the samples between domains, in order to achieve more relevant answers to the questions, we got much smaller samples. It is a pity that the generally largest of the four written domains, NA, is without much interest, since it now only makes the samples from the other domains smaller, but this must be accepted. Here are some examples of Written sample sizes:

(4) *So* Written Fiction 75, *therefore* Written Academic 107

(5) *So* Written: Newsp 20 and Academic 19

(6) *Therefore* Written: Fiction 4 and Newsp 6

The uncertainty not only depends on these sample sizes but also on the population frequency of the characteristic studied, which could for example be the frequency of a specific SoC characteristic or a combination of Rel and SoC. The standard error is highest when the frequency is $1/2$ ($= 50\%$), when the standard error is $\frac{1}{2\sqrt{n}}$, where n is the domain sample size. In the examples above, the standard errors in this 'worst case' scenario are

(4) 0.05 – 0.06 (5–6 percentage units);

(5) ≈ 0.12;

(6) 0.20–0.25.

In case **(6)** the standard error is a too simple measure of uncertainty, however, but it is anyhow clear that the uncertainty is huge. When the frequency in question is 0.10 (10%) or 0.90, instead of $\frac{1}{2}$, the corresponding standard errors are smaller by a factor of 0.6:

(4) 0.03 – 0.04;

(5) ≈ 0.07;

(6) 0.12–0.15.

Here are a couple of illustrations:

(7) *Therefore* Written Academic, n=107: SoC=Author, 60 instances of 107 => proportion 0.56 ± 0.05 (more precisely ± 0.048)

(8) *Therefore* Written Academic, n=107: Modal (type 1 or 2) 8+27=35 instances => proportion 0.33 ± 0.05 (± 0.046);
Note: Nonmodal is complementary, thus 0.67 ± 0.05 (same *s.e.*)

Many combinations of Rel and SoC has a zero number of observed instances. The population proportion may of course be zero (and in some cases it may be natural to assume it is zero), but at the other end, how large can the population proportion realistically be? This depends on the sample size, and let us consider a couple of examples. First, we return to example **(8)** above, with $n = 107$. Suppose a zero is observed (e.g. the combination $\{SA, Ch\}$), and that such a zero was not a very unlikely event to happen, more precisely that the probability of a zero was at least 5%. The demand on the population for this to happen is a population proportion of at most 0.012, i.e 1.2%. Thus we can feel pretty sure the *Therefore* Academic corpus domain did not have much more than 1% instances of the combination $\{SA, Ch\}$. A general formula for small proportions is $\ln(20)/n$, where ln is the so-called natural logarithm function.
Let us now go to example **(6)**, the small *Therefore* Newspaper domain, which has also a zero for $\{SA, Ch\}$, but with a domain sample size of only $n = 6$. Then the domain could of course have much more of $\{SA, Ch\}$) without any of them seen in the small sample. In this case the demand on the domain to make this plausible is that the domain population proportion of $\{SA, Ch\}$ be at most about 40% (that is, it could be quite large).

Note that the 'small' samples for AAR and FTR Spoken are not small as explorations of the corpus. They comprise the whole corpus populations of these connectives. On the other hand we know from these population sizes that AAR and FTR are quite rare in Spoken language, and as a consequence that the BNC is insufficient, if we are interested in the (rare) use of them.

V.3. Uncertainty in the 'probabilistic approach', whole domains

The simplest situation for the probabilistic approach is when we consider the choice of connective within a certain domain, without specifying Rel, SoC or Modal. For each connective, the domain population sizes are known, as soon as we know the target proportions, because the crude domain population sizes are known (without statistical uncertainty). if we can calculate the domain size for each connective, we immediately also know the probability that a random choice of instance would result in a specific connective. It is just to sum the total number of target instances over the four connectives, and see how large proportion of this sum was taken by each connective. Thus, except for the need for the target proportions, discussed in Section 1, there is no statistics involved. As remarked in Section 1, the sampling errors in the target proportions are small, although for connective *so* there is a variation between domains to consider and adjust for.

V.4. Uncertainty in the 'probabilistic approach', domain parts

When we further specify a category of Rel, SoC, or Modal, or a combination of such categories, the situation gets more complicated. The domain population size of that category must be estimated from the sample data, for each connective separately. We consider two examples jointly, the (Rel, SoC) combinations $\{SA, A\}$ and $\{E, A\}$ in the Written Academic domain (cf. Example **(7)** above. We neglect connectives AAR and FTR, which are rarely chosen by the language user and therefore have little *uncertainty* effect on *so* and *therefore*. Also, as we will see, it is enough complicated to make statements of uncertainty about the choice between two connectives. For *so*, the domain sample size is n=19, one instance of $\{SA, A\}$ and 8 of $\{E, A\}$. For *therefore*, the corresponding domain sample size is $n = 107$, with 3 instances of $\{SA, A\}$ and 55 of $\{E, A\}$. In the $\{SA, A\}$ category, the probability for choice of *so* is estimated to be 40%, and for choice of *therefore* 59%. In the $\{E, A\}$ category, the corresponding probabilities are 22% and 74%. After elimination of AAR and FTR this probability increases to 23% and 77%, respectively.

As in the previous section, if we had known the total number of instances of *so* and *therefore* in the corpus that satisfy the joint specification of domain, Rel and SoC, it would just be to count the proportion of *so* and *therefore* respectively in their total (sum over *so* and *therefore*). The total

number per connective satisfying the specification is not known, but it is naturally estimated, by multiplying the domain total with the domain sample relative frequency of instances satisfying the specification of Rel and Soc. From the previous paragraph we have for example the relative frequencies 1/19 and 8/19 for *so* ($\{SA, A\}$) and $\{E, A\}$, respectively) and the relative frequencies 3/107 and 55/107 for *therefore*. We see again the problems discussed in Section 0.1. The sampling uncertainty for *therefore* is small, but for *so* it is quite large in this case.

Instead of treating the proportions of *so* and *therefore* (the ratios to their sum), that is, their respective estimated probabilities of choice, as described above, we go over to their odds, which has mathematical/statistical advantages. More precisely. the odds for *so* is the ratio of the probabilities of choice of *so* and of 'not-so', (that is of *therefore*). From the odds, we can calculate the probabilities (and vice versa): add 1 to the odds, invert, and subtract from 1 (so odds = 1 corresponds to equal chances, probability = 1/2). The odds for category $\{SA, A\}$ is about 4 to 6, more precisely estimated as 0.69. For $\{E, A\}$ it is estimated to be 0.30 The next step is to see if one of the two contributions to the uncertainty dominates over the other, that is if the uncertainty coming from *so* dominates over that coming from *therefore*, or vice versa, or if they contribute equally. Their relative contributions of variance is estimated by $1/y - 1/n$, where y is the number of instances observed in the sample of size n (specified Rel, SoC, domain, connective; formula derivation excluded). In the two examples we get:

$\{SA, A\}$ Estimated variance contribution from *so* $= 1 - 1/19 = 0.95$;
from *therefore* it is only $1/3 - 1/107 = 0.32$;

$\{E, A\}$ Estimated variance contribution from *so* $= 1/8 - 1/19 = 0.07$;
from *therefore* it is only $1/55 - 1/107 = 0.009$.

We note that for each (Rel, SoC) combination the contribution from the small domain (for *so*) is substantially higher than that from the relatively large domain (for *therefore*). This is not surprising. We first consider the second (Rel, SoC) combination.

For the $\{E, A\}$ category of the Academic domain, with odds ratio 0.30, we have seen that the contribution to uncertainty from *so* dominates. The relative variance corresponds to a relative standard error of $\sqrt{0.07} \approx 0.3$. This yields an interval around odds 0.3 by adding and subtracting about 30% of the odds ratio value itself, which yields the odds ratio interval 0.2 to 0.4 . Transforming back to probabilities of choice we get the interval 0.18 to 0.28 around the estimate of 0.23 for choice of *so*, and the complementary

values for *therefore*. This is a quite reasonable precision to draw conclusions from.

For the $\{SA, A\}$ category of the Academic domain, the situation is more difficult, and the uncertainty in the odds ratio will be much higher. The standard error calculus is not sufficient, cf. the calculation for zero instances toward the end of Section 0.1, related to example **(6)**.

The statistical error in *so* dominates over the error in *therefore*, so we only consider the former. One instance of $\{SA, A\}$ in a domain sample of 19 tells that the probability for choice of *so* can be neither very close to zero nor very high, so we can construct a two-sided confidence interval for the probability by exact calculations in the binomial distribution. We have chosen the 70% confidence level in order to match the $\pm\, s.e.$ interval. The confidence interval in this case was (0.0085, 0.167). This was used to get an interval for the odds of *so* versus *therefore*, which was finally transformed to an interval for the probability of *so* (or of *therefore*): $0.10 \leq \Pr(so) \leq 0.68$.

With a zero number of instances of *so* (or *therefore*), as for example in Academic domain $\{E, CS\}$, a one-sided confidence interval would have been necessary (and natural). In this case, more care is needed in the calculation of an interval for the odds, since when zero instances are observed, we are comparing with zero (or infinite) estimated odds. Details are omitted.

The uncertainty in choice probabilities is to a large extent controlled by the number of instances of *so* for the discourse context in question. If this number is not very small, a large number of *so* is expected in the corpus, and the probability of choice of *so* cannot be small. When the number of instances of *so* is small, or even zero, the probability of choice of *so* is likely to be quite uncertain. On the other hand, for AAR or FTR the probability of choice will be quite small for almost all discourse contexts, and so will the uncertainty in absolute terms for AAR and FTR, whereas their relative uncertainty is likely to be large.