UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Swansea University
Prifysgol Abertawe

Stimuli design: Frequency measure challenges and the perennial problem of L2 learner input

Rachel Klassen[1] & Vivienne Rogers[2]

[1]*LAVA & AcqVA*, UiT The Arctic University of Norway
[2]Swansea University

Funded by the European Union

AcqVA
Acquisition Variation Attrition

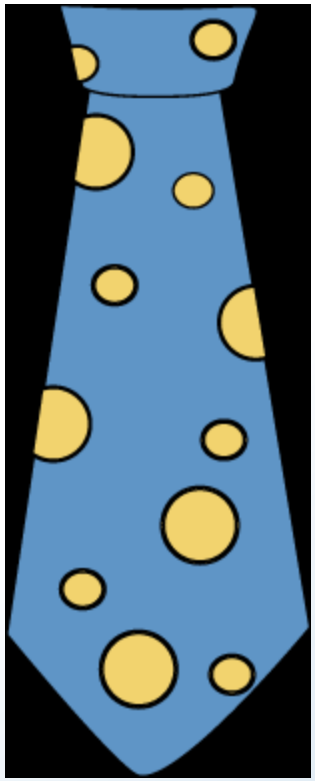
UNIVERSITY OF TROMSØ

Why consider word frequency?

- Frequency of words affects language processing (e.g., Hopp 2015)
- High frequency words produced and comprehended more quickly than low frequency ones
- L1 word frequencies affect responses in L2 tasks (e.g., Lemhöfer et al 2008)

Word frequency varies across languages

- *Une cravat* = 4000k band in French



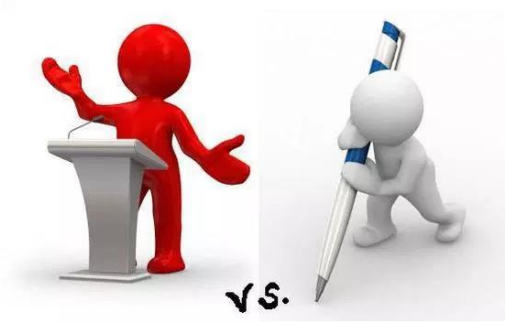
- *A tie* = 2000k band in English

+ cultural considerations in L2/Ln vocabulary

Research questions

What type of corpora should be used to measure word frequency?

How can word frequency be balanced cross-linguistically?



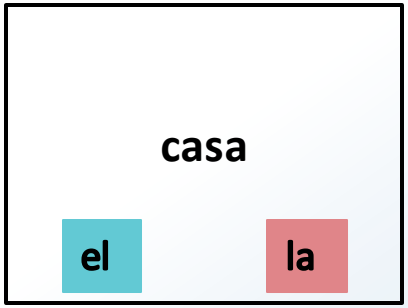
Why was this an issue?

Four different L1s

Learning Spanish as an L2/Ln

What we had to consider

- **Task: L2 Spanish gender decision task**
  - participants presented with individual nouns in Spanish
  - press a button to indicate the correct definite determiner (el_M/la_F)
  - reaction times and accuracy measured



- **Participants**
  - intermediate-advanced level of proficiency in L2 Spanish
  - four different L1 groups (Norwegian, German, Dutch, Latvian)

What we had to consider

- **Experimental conditions**
  - according to the gender of the noun in the L1 and the L2
  - *main conditions:* same gender (congruent), different gender (incongruent)
  - *subconditions:* broken down according to the specific gender value

	Congruent		Incongruent		L1 Neuter	
	L1	*L2*	*L1*	*L2*	*L1*	*L2*
subconditions	masc fem	masc fem	masc fem	fem masc	neut neut	masc fem

- **Stimuli**
  - avoided cognates and nouns that have multiple translation equivalents (e.g., *Die Linse* is lens and lentil)
  - controlled for frequency, number of letters, percent overlap between L1 and L2 nouns

What we had to consider

- **Additional considerations**
  - transparency of the gender marking on Spanish nouns
  - other gendered languages known to entire participant pool (e.g., Russian in Latvia)
  - cross-linguistic differences in how gender is marked (e.g., definiteness in Norwegian)
  - language variation in L1 and L2 (e.g., Peninsular vs Latin American Spanish; Norwegian dialects, changes to gender system)

www.sketchengine.eu

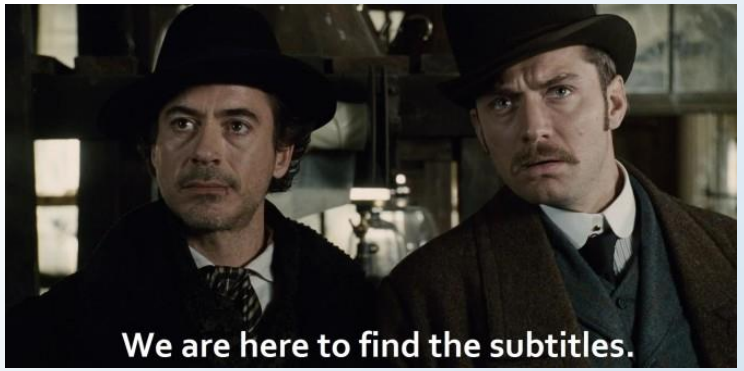
German	Araneum Germanicum Maius [2013]	875,465,845	
German	CHILDES German Corpus	5,941,266	
German	DGT, German	45,380,666	
German	EUR-Lex German 2/2016 (old WS grammar)	529,807,527	
German	EUR-Lex judgments German 12/2016	35,297,517	
German	EUROPARL7, German	47,805,055	
German	German Corpus for SkELL 1.0	769,810,745	
German	German Web (deWaC)	1,348,188,416	
German	**German Web 2013 (deTenTen13)**	**16,526,335,416**	
German	German Web 2013 sample (deTenTen13)	54,615,446	
German	GerManC (German Newspapers 1650-1800)	667,310	
German	OPUS2 German	125,229,773	
German	Parsed German Web (sDeWaC)	755,165,551	
German	Timestamped JSI web corpus 2014-2016 German	1,987,759,563	
German	Timestamped JSI web corpus 2014-2018 German	3,152,046,174	
German	Timestamped JSI web corpus 2018-03 German	90,796,531	
German	Timestamped JSI web corpus 2018-04 German	84,762,551	

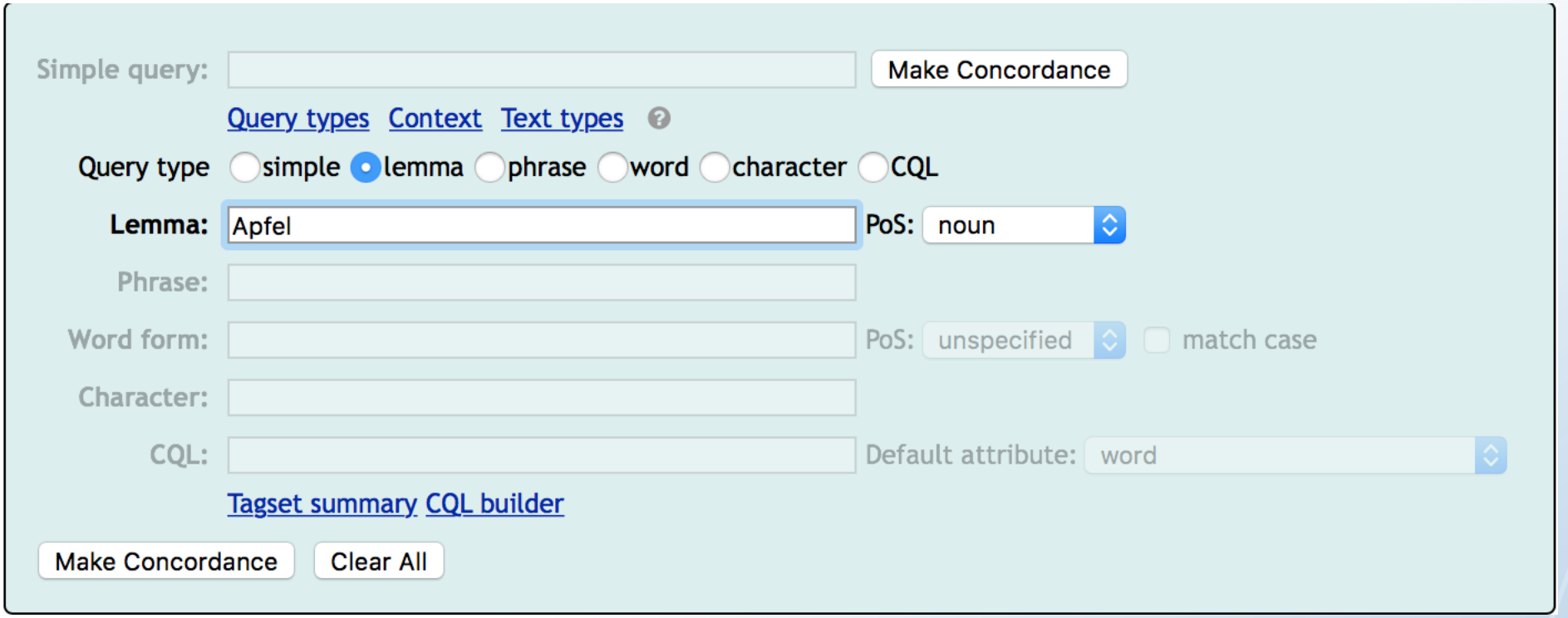
Jakubíček et al (2013)

- TenTen web corpus
- 30+ languages

Tiedemann (2016)

- OPUS2 subtitling corpus: OpenSubtitles 2011
- 54 languages




Simple query:
Make Concordance
Query types
Context
Text types
Query type
simple
lemma
phrase
word
character
CQL
Lemma:
Apfel
PoS:
noun
Phrase:
Word form:
PoS:
unspecified
match case
Character:
CQL:
Default attribute:
word
Tagset summary
CQL builder
Make Concordance
Clear All


Sketch Engine
German Web 2013 (deTenTen13)
Search in detenten13_rft3 (trial)
Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Corpus info
My jobs
User guide
Save
Make subcorpus
View options
KWIC
Sentence
Sort
Left
Right
Node
References
Shuffle
Sample
Filter
Query (Apfel)-n 311,928 (15.70 per million)
Page 1 of 15,597 Go Next | Last
allergie-i... zu erklären: Ich bringe gerne das Beispiel des Apfels , weil hierdurch die Funktionsweise am
guerilla-m... möglichst schnelle Bekanntheit gesteckt. Der Apfel mußte dazu wohl oder übel selbst auf die Straße.
guerilla-m... Fräulein, darf ich´s wagen, Ihnen diesen Apfel anzutragen? " zumindest kurzzeitig den Zwang
guerilla-m... den Zwang verspürt haben, sich nach Prinz und Apfel umzudrehen... Wie viel bekanntheits- und
sylbach.de... gefeiert: An kleinen Tischen wird man Brot, Äpfel und Wasser miteinander teilen. Die Lippische
neuoffenba... - Paradiese und kann sich an dem dargereichten Apfel nimmer satt fressen! [BM.01_008,15] O große
gartenrund... . Es gibt eine große Sortenvielfalt an Äpfeln (alte und neue Sorten), Zwetschgen,
schalom-ra... auf dem Flügel Maria Boguslavskaja. Wein, Apfel mit Honig, Gebäck erfreuten Gäste und
schalom-ra... feiern und das neue Jahr 5772 zu begrüßen. Wein, Äpfel mit Honig, Gebäck, Feigen und nicht zuletzt ein
schalom-ra... Wein, traditionelle Hefezöpfe sowie Äpfel , die in Honig getunkt werden, um ein "süßes"
esotericon... , Unglück auf Reisen... 1. Halbiere einen Apfel und reibe die eine Hälfte mit (möglichst
esotericon... aus, was gebannt werden soll. Danach wird der Apfel mit einem Zahnstocher oder einem anderen
esotericon... wieder zusammengefügt. Umwickle dann den Apfel mit einem grünen Band und vergrabe ihn an einer
esotericon... einer Stelle, die nur dir bekannt ist. Wenn der Apfel verfault ist, hat sich das Problem erledigt! 2.
gartenforu... <gap/> <gap/> Gerade bei Obstgehölzen wie z.B. Äpfel , Birnen oder Kirschen gibt es oft viele Frage,
orthodoxia... von den verschiedenen Früchten und Beeren (Apfel -, Birnen-, Kirsch-, Schlehdorn-, Himbeer
km-bw.de aus Deutschland gegenüberstellt, vergleicht Äpfel mit Birnen", sagte Rau. Die
mahnung-ge... kaum möglich sein, auch wenn Parteichef Holger Apfel zuletzt auf einer Pressekonferenz in Pampow
mahnung-ge... der sächsische NPD-Fraktionschef Holger Apfel demonstrativ gemäßigt auf und nennt dieses
mahnung-ge... Reden Bezug auf demokratische Politiker. Apfel hat sich von den NSU-Morden distanziert -
Page 1 of 15,597 Go Next | Last

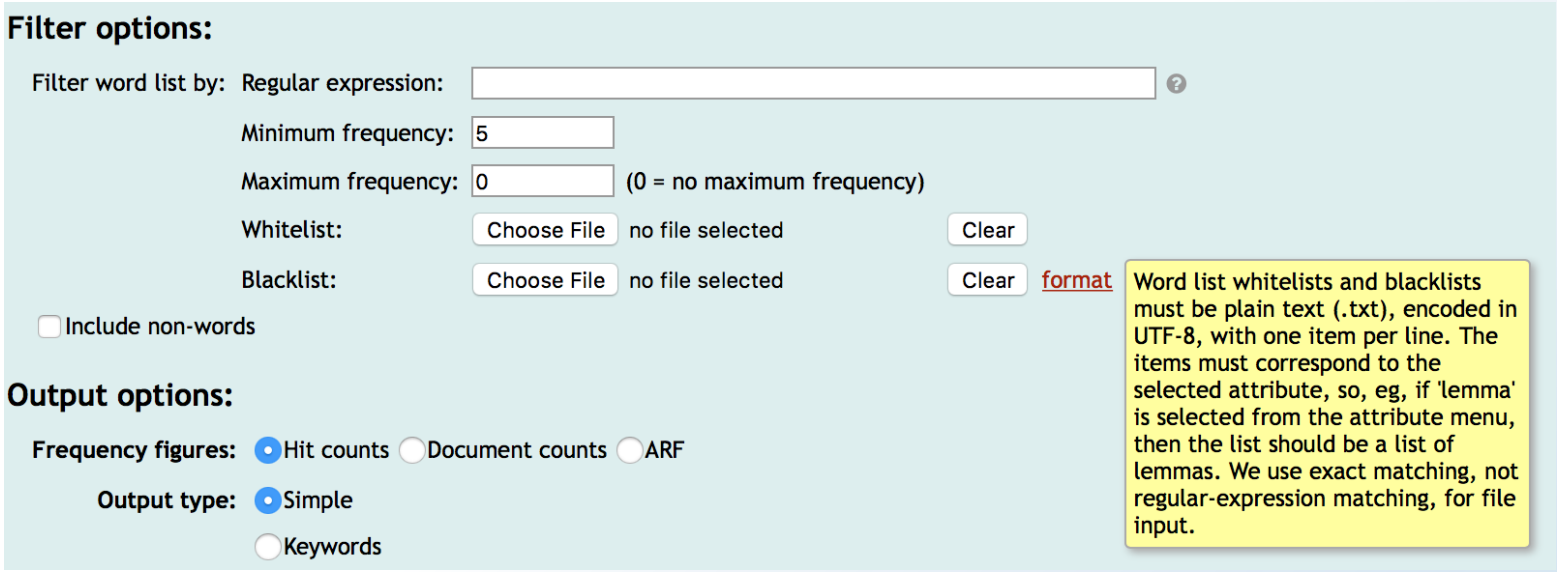
Automating the search

- Excel (csv) file for each language's target words
  - Norwegian
  - German
  - Dutch
  - Latvian
  - Spanish
- C. 250 words per language per corpus
- Over 2500 searches.
- API (application programme interface)
- Submit query to wsketch (word sketch)
- Each language/ corpus searched separately
- Output to excel
- Capped 900 searches per hour
  - 5 sec pause built into programme = 650 ph

```
data['corpname'] = options.corpname
data['usesubcorp'] = options.usesubcorp
data['lpos'] = options.pos

wordReader = csv.reader(sys.stdin, delimiter='\t')
spamWriter = csv.writer(sys.stdout, delimiter='\t',quotechar='"', quoting=csv.QUOTE_ALL)

for row in wordReader:
  data['lemma'] = row[0]
  data['lpos'] = options.pos
  if len(row) > 1:
    if row[1].upper()[0] == 'V':
      data['lpos'] = '-v'
  else:
    row.append('N')
  d = requests.get(base_url + '/wsketch', params=data).json()
  #d = requests.get(base_url + '/corp_info', params=data).json()
  pp = pprint.PrettyPrinter(indent=2)
  #pp.pprint(d)
  #pp.pprint(data)
  if (options.lookup_pos):
    mylposdict = d.get('lpos_dict',{data['lpos']: data['lpos']})
  else:
    mylposdict = {'adjective': '-j', 'adverb': '-r', 'noun': '-n', 'verb': '-v'}
  mylposkeylist = list(mylposdict.keys())
  mylposvaluelist = list(mylposdict.values())
  mylist = [d.get('freq',0), d.get('relfreq',0), d.get('lemma',row[0]), row[1], mylposkeylist[mylposvaluelist.index(d.get('lpos',data['lpos']))],
d.get('corp_full_name',data['corpname']),d.get('usesubcorp',data['usesubcorp'])]

  spamWriter.writerow(mylist)
  time.sleep(5)
```

Other options (for non programmers)

- Sketch Engine has Whitelist option
- Needs to be in txt format and lemmatised



Working with the stimuli

- **Step 1: Confirm translation equivalents**
  - consult with speaker(s) who have the same profile as target group
  - high-proficiency speakers can give more detailed information
  - speakers with the same proficiency as target group allow for the elimination of unknown L2 nouns

- **Step 1 tasks: L1 translation & Existing list modification**

Por favor, traduce las palabras en español al noruego e indica el género de la palabra en noruego. Es importante poner la primera palabra que se te ocurre y de no usar el diccionario. Si no conoces la palabra en español, indica eso simplemente en la columna 'noruego'.

español	noruego	género (noruego)
abanico	-	
abrigo	yttertøy	n
aeropuerto	flyplass	m
agua	vann	n
ajo	hvitløk	m
algodón	bomull	m
almohada	pute	f

Working with the stimuli

- **Step 1 tasks: Existing list modification**

	A	B	C	D
1	**Spanish**	**Norwegian**	**Comments**	**Amigos falsos**
2	abanico	vifte		
3	abrigo	frakk	Aquí la palabra que se me ocurrió primero fue yttertøy, pero es imprecisa. Frakk es una palabra buena, y es lo que usan los hombres. Por tanto, según yo, también puede ser kåpe, que sería una prenda más para mujer.	
4	bandera	flagg		
5	barco	båt		bark = corteza de árbol
6	bolsa	pose		
7	bolso	veske		
8	bosque	skog		buske = arbusto
9	botella	flaske		
10	brazo	arm		
11	bufanda	skjerf		
12	cabaña	hytte		
13	cabeza	hode		
14	calcetín	sokk		
15	calle	gate		kall = llamamiento (vocación religiosa). kalde = fríos (plural, específico del adjetivo)
16	calma	ro		
17	cama	seng		kam = peine
18	camino	vei		
19	camión	lastebil		kameleon = camaleón
20	camisa	skjorte		
21	camiseta	trøye	Aquí también se usa a menudo “t-skjorte”	
22	campana	klokke		
23	canción	sang		
24	cara	ansikt		kar = hombre
25	cárcel	fengsel		karse = berro
26	carne	kjøtt		
27	casco	hjelm	Casco también puede significar cáscara (por ejemplo de las nueces)	kasko = un tipo de seguro para autos
28	cebolla	løk		

Working with the stimuli

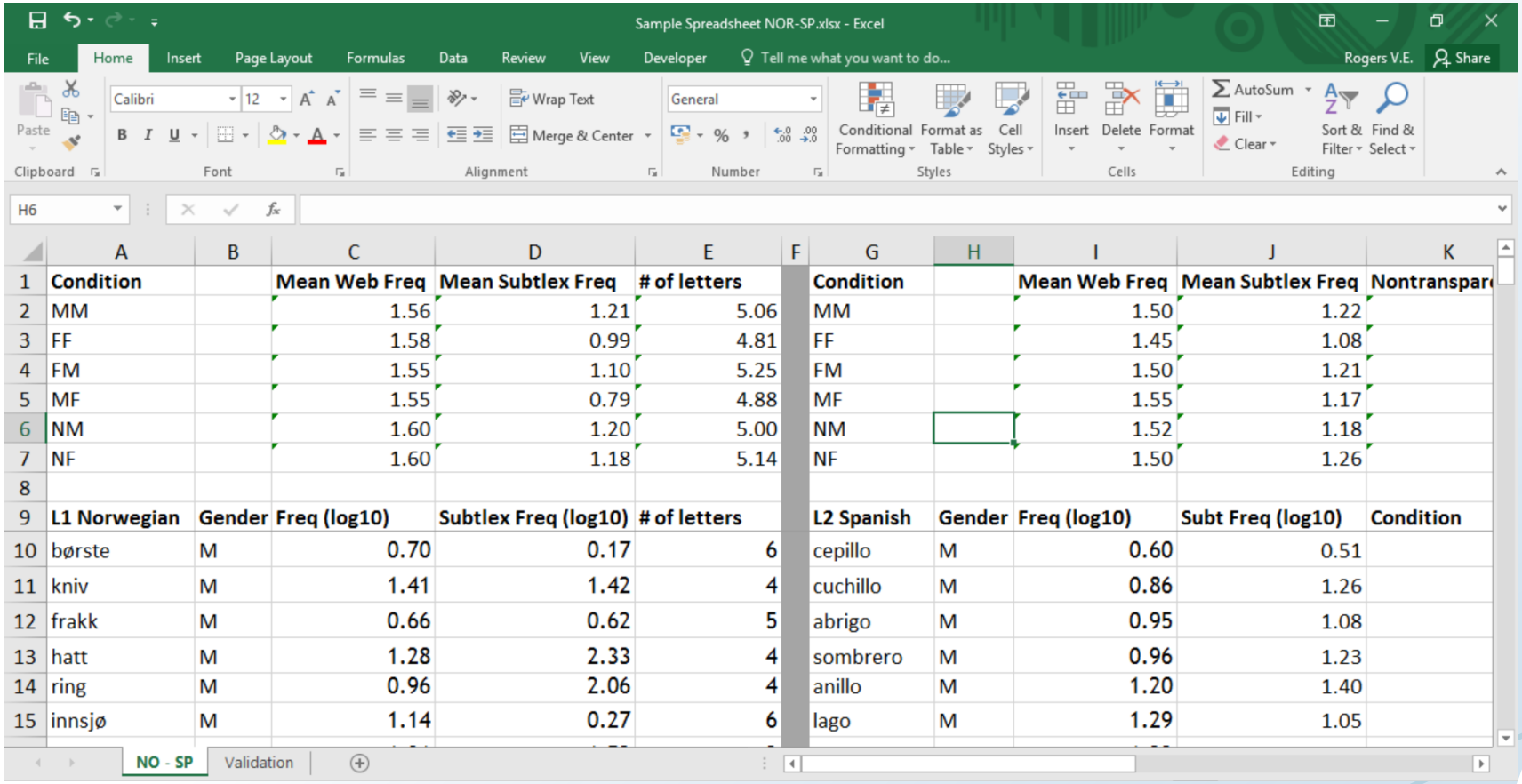
- **Step 2: Add frequencies, relevant coding & pare down list**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**L1 Norwegian**	**Gender**	**Freq (per mill)**	**Freq (log10)**	**Subt Freq (per mill)**	**Subt Freq (log10)**		**L2 Spanish**	**Gender**	**Freq (per mill)**	**Freq (log10)**	**Subt Freq (per mill)**	**Subt Freq (log10)**	**Condition**
2	himmel	M	53,60	1,73	12,85	1,11		cielo	M	42,40	1,63	55,2	1,74	1
3	hvete	M	8,12	0,91	1,66	0,22		trigo	M	8,80	0,94	1,9	0,28	1
4	hvitløk	M	14,70	1,17	1,78	0,25		ajo	M	16,00	1,20	2,2	0,34	1
5	jus	M	4,20	0,62	2,61	0,42		zumo	M	11,50	1,06	2	0,30	1
6	penge	M	301,90	2,48	42,69	1,63		dinero	M	173,50	2,24	234,8	2,37	1
7	verdi	M	136,30	2,13	8,61	0,94		valor	M	277,90	2,44	24,8	1,39	1
8	agurk	M	5,30	0,72	1,06	0,03		pepino	M	2,80	0,45	2,5	0,40	1
9	arm	M	62,06	1,79	14,62	1,16		brazo	M	36,50	1,56	38,6	1,59	1
10	bil	M	385,60	2,59	136,35	2,13		coche	M	171,20	2,23	72,4	1,86	1
11	blyant	M	9,90	1,00	2,98	0,47		lápiz	M	6,20	0,79	5,1	0,71	1
12	børste	M	5,00	0,70	1,47	0,17		cepillo	M	4,00	0,60	3,2	0,51	1
13	båt	M	133,00	2,12	31,74	1,50		barco	M	42,20	1,63	42,2	1,63	1
14	dag	M	2015,50	3,30	756,70	2,88		día	M	1114,30	3,05	355,6	2,55	1
15	dal	M	21,90	1,34	5,10	0,71		valle	M	41,00	1,61	5,6	0,75	1
16	dessert	M	15,20	1,18	9,33	0,97		postre	M	12,60	1,10	5,5	0,74	1
17	farge	M	207,30	2,32	17,23	1,24		color	M	183,30	2,26	27,1	1,43	1
18	feil	M	177,80	2,25	308,00	2,49		error	M	105,80	2,02	52,7	1,72	1
19	fersken	M	2,60	0,41	4,16	0,62		melocotón	M	2,40	0,38	1,1	0,04	1
20	flyplass	M	47,10	1,67	7,67	0,88		aeropuerto	M	63,60	1,80	14,3	1,16	1
21	frakk	M	4,60	0,66	4,16	0,62		abrigo	M	8,90	0,95	12	1,08	1
22	gaffel	M	5,40	0,73	2,38	0,38		tenedor	M	2,70	0,43	6,8	0,83	1
23	genser	M	18,80	1,27	3,74	0,57		jersey	M	6,50	0,81	2,9	0,46	1
24	hage	M	73,30	1,87	7,71	0,89		jardín	M	49,00	1,69	14,5	1,16	1
25	hammer	M	8,44	0,93	6,91	0,84		martillo	M	3,50	0,54	4,5	0,65	1
26	hatt	M	18,90	1,28	214,90	2,33		sombrero	M	9,10	0,96	17,1	1,23	1
27	heis	M	12,50	1,10	7,78	0,89		ascensor	M	18,10	1,26	8,5	0,93	1
28	hjelm	M	11,70	1,07	4,53	0,66		casco	M	31,40	1,50	7,3	0,86	1

NO - SP | Validation | +

Working with the stimuli

- **Step 3: Balance mean frequencies by condition and language**



	A	B	C	D	E	F	G	H	I	J	K
1	**Condition**		**Mean Web Freq**	**Mean Subtlex Freq**	**# of letters**		**Condition**		**Mean Web Freq**	**Mean Subtlex Freq**	**Nontranspar**
2	MM		1.56	1.21	5.06		MM		1.50	1.22	
3	FF		1.58	0.99	4.81		FF		1.45	1.08	
4	FM		1.55	1.10	5.25		FM		1.50	1.21	
5	MF		1.55	0.79	4.88		MF		1.55	1.17	
6	NM		1.60	1.20	5.00		NM		1.52	1.18	
7	NF		1.60	1.18	5.14		NF		1.50	1.26	
8											
9	**L1 Norwegian**	**Gender**	**Freq (log10)**	**Subtlex Freq (log10)**	**# of letters**		**L2 Spanish**	**Gender**	**Freq (log10)**	**Subt Freq (log10)**	**Condition**
10	børste	M	0.70	0.17	6		cepillo	M	0.60	0.51	
11	kniv	M	1.41	1.42	4		cuchillo	M	0.86	1.26	
12	frakk	M	0.66	0.62	5		abrigo	M	0.95	1.08	
13	hatt	M	1.28	2.33	4		sombrero	M	0.96	1.23	
14	ring	M	0.96	2.06	4		anillo	M	1.20	1.40	
15	innsjø	M	1.14	0.27	6		lago	M	1.29	1.05	

Working with the stimuli

- **Step 3: Things that helped us**
  - start with the most restricted condition and match other ones to it (i.e., in terms of frequency and number of stimuli)
  - move eliminated stimuli to another list rather than just deleting them and make a note as to what the issue was (e.g., false cognate, unknown to L2 speakers)
  - allow more time than you expect
    - 'double your estimate and increase by an order of magnitude' ☺

	A	B	C	D	E	F	G	H	I	J	K
1	**Condition**		**Mean Web Freq**	**Mean Subtlex Freq**	**# of letters**		**Condition**		**Mean Web Freq**	**Mean Subtlex Freq**	**Nontranspar**
2	MM		1.56	1.21	5.06		MM		1.50	1.22	
3	FF		1.58	0.99	4.81		FF		1.45	1.08	
4	FM		1.55	1.10	5.25		FM		1.50	1.21	
5	MF		1.55	0.79	4.88		MF		1.55	1.17	
6	NM		1.60	1.20	5.00		NM		1.52	1.18	
7	NF		1.60	1.18	5.14		NF		1.50	1.26	

Problems / Limitations

- Differences in frequency cross-linguistically
- Differences in frequency between two corpora

L1 Norwegian	Gender	Freq (log10)	Subtlex Freq (log10)
kjøleskap	Nt	1.41	0.33

Common words that most learners know not in corpora/very low frequency

- Checked learner dictionary (Davies 2006) but it's not based on learner data but written corpora
- Check learner corpora?
  - Checked talkbank.org but either elicited or small sample
- Check exam boards?
  - Wide variation across countries
  - CEFR lists?

Conclusions

- Using the same protocol allows for consistency across languages…
  - within the same experiment
  - with different participant groups

- Using both web and subtitling corpora…
  - offers complementary measures of language in general
  - gives more representative sample of L2 learner input

- This particular approach is applicable to many different languages

UiT

THE ARCTIC UNIVERSITY OF NORWAY

Thank you!

Swansea University
Prifysgol Abertawe

Rachel Klassen
rachel.klassen@uit.no

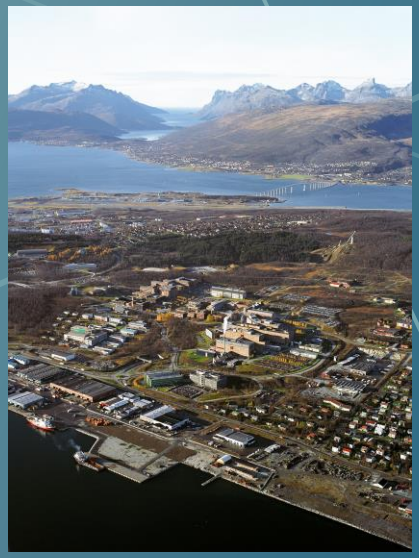
GenBiLex
http://site.uit.no/lava/genbilex/

LAVA
http://site.uit.no/lava/

AcqVA
https://site.uit.no/acqva/

Vivienne Rogers
v.e.rogers@swansea.ac.uk

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 748966.



UNIVERSITY OF TROMSØ