to appear in: Gotzner, Nicole and Uli Sauerland (ed.). 2021. "Measurements, Numerals, and Scales: Essays in Honour of Stephanie Solt." Palgrave. doi: 10.1007/978-3-030-73323-0

Quantifying the register of German quantificational expressions: A corpus based study*

Uli Sauerland

Abstract Numerical and quantificational expressions differ not only in their logical meaning, but also in their conditions of use. In particular, One hypothesis suggested in previous work states that vague expressions should be associated with informality and precise expressions with formality. We introduce a novel corpus measure, the SOLT, to quantify formality of an expression and show that it correlates well with established data on formality. We then apply the SOLT to numerical and quantificational expression to test the hypothesis that vagueness and informality are related. We find that there is some support for the hypothesis, but the relation is not straightforward as claimed.

Numerical and quantificational expressions are well studied and understood in many dimensions relevant to their semantics and pragmatics, in particular, their truth-conditional thresholds, their polarity properties, and their roundness (Solt 2006; 2009; 2014; 2015a; Solt et al. 2017 and others). But they have not been investigate much from the perspective of register.[1] Such an investigation promises to contribute both our understanding of the pragmatics of quantifiers and the establishment of a link between sociolinguistic and grammatical theories of meaning (Burnett 2017; Smith et al. 2010, and others). This paper presents a corpus based measure of colloquiality, the SOLT, which is a log-ratio of the written and spoken frequency based on publicly available text corpora in Section 1. I also validate the SOLT as a quantitative measure correlating with dictionary writers intuitions about style or register.

After having established the SOLT, I apply it to German quantificational determiners in Section 2. In particular, I seek to test the hypothesis that vague (or imprecise) quantifiers generally have a less formal register than precise quantifiers. The hypothesis is at least suggested by the work of Brown & Yule (1983) on differences between spoken and written language and that of Coupland (2007) between mundane and high performance. I show that the data from cardinals don't corroborate the hypothesis. I furthermore argue that in the domain of indefinites, I observe that ignorance marking among singular indefinites correlates with register, while among plural indefinites register variation appears unrelated to meaning differences. Overall I conclude that the SOLT is useful and provides some support for the hypothesis, but that additional sources of evidence need to be explored.

* I am grateful to Stephanie Solt for many inspiring and motivating remarks on this and other work of mine. I also thank Artemis Alexiadou and Roland Mühlenbernd for helpful comments, Kai Sauerland for providing a Python skript, and Renata Shamsutdinova for editorial help. This paper derives from the work in project A05 of CRC 1412 and I am grateful for the financial support of the DFG.

[1] Duffley & Larrivée (2012); Solt & Stevens (2018) and Stevens & Solt (2018) investigate English *some*.

1 The SOLT

Though register is not reducible to the difference between literal and oral language, register nevertheless correlates with literality. I therefore propose a measure of literality for German words and then investigate to what extent it provides similar results as other sources of register information. I call the measure I propose the SOLT (for *Schriftsprache/Oralsprache logarithmisch transformiert*, 'written language/oral language logarithmically transformed'), defined for a German dictionary lemma *ell* as the binary logarithm of the ratio of the frequency of *ell* in a corpus of written language over the frequency of *ell* in a corpus of spoken language (see (1)). The SOLT is inspired by the notion log-likelihood ratios in other areas.

$$\text{SOLT}(\ell) = \log_2 \frac{\text{frequency(written)}(\ell)}{\text{frequency(spoken)}(\ell)} \tag{1}$$

I decided to use the binary logarithm because the resulting figures have more intuitive appeal than the natural or decimal logarithm: A SOLT value of 1 means a lemma occurs twice as often in the written as in the oral corpus, while a SOLT of -2 means it is a quarter as frequent in writing as it is in speech. Consider three German verbs expressing *sell* as an example: *verticken*, *verkaufen*, and *veräußern*. Their SOLT values and the underlying occurrence counts are shown below. The SOLT corresponds to native speaker intuitions that there are register differences between the three.

lemma	written occ.	oral occ.	total occ.	SOLT
verticken	4	125	129	-5.65
verkaufen	5094	10948	16042	-1.79
veräußern	194	22	216	2.46

The source of corpus information I use is the DWDS (www.dwds.de), and specifically, the *kern* ('core') corpus as the source of written language and the *untertitel* ('subtitle') corpus for spoken language. Both of the choices come with several confounds that I mention below, but the practical reason that both are of substantial size with 121 million and 76 million words respectively and are freely and easily accessible trumps these other considerations.

1.1 Methods for the Independent Validation of the SOLT

Does the SOLT correlate with register or style like we saw in the above example of verbs of selling? In this subsection, I report a brief study that validates the SOLT. I use the *Duden* (Dudenredaktion 2015) as an independent source of register information for lexical items. The Duden offers some information about register as evidenced by the table of abbreviations. Duden uses about 400 abbreviations, and I estimate that about one third of the abbreviations introduced relate to specific circumstances of use of a word. In the introduction of the terms referring to style on p. 18 the authors seem to indicate a scale of unmarked, colloquial, harsh, and vulgar, and mention elevated and specialized language as opposite deviations from the unmarked.

> *Was manchen Benutzern normalsprachlich […] erscheint, ist für andere schon »ugs.« (= umgangssprachlich), ja gar »derb« oder sogar »vulg.« (= vulgär). Ähnlich verhält es sich mit Bewertungen wie »geh.« (= gehoben) oder »fachspr.« (= fachsprachlich).*
>
> *(What some users judge to be 'normal language' is for other already 'colloquial', or even 'harsh' or even plain 'vulgar'. It is similar with judgments like 'elevated' or 'specialized'.)*

In the following study, I focus on the abbreviations for 'colloquial' (*ugs.*), 'elevated' (*geh.*) and 'specialist language' (*fachspr.*) as indicators of register because these three were most frequent and easily findable in an electronic version of the dictionary.[2] For the adjective *derb* ('harsh'), I couldn't reliably determine its frequency: the string 'derb' occurs 831 times, but that includes occurrences within a word such as in *Rinderbraten* ('beef roast'). The annotation '(derb)' only occurs 116 times. Several of the other abbreviations introduced occur only rarely in the dictionary. For example, there are only 24 occurrences of *vulg.* ('vulgar'), 8 of *amtl.* ('governmental'), and 23 of *standardsprachl.* ('belonging to the standard language'). The abbreviation *geb.* has 428 occurrences, but is used to abbreviate both *gebildet* ('educated') and *geboren* ('born'), so I also decided not use it. The Duden also includes abbreviations for the about 240 specialized registers listed on page 21 such as *Bergmannsprache* ('miners language') , *Kirchensprache* ('church language'), and *Zollwesen* ('customs affairs'). I do not consider these since their frequency might interact with the popularity of specific subjects in written texts vs. movies. The three abbreviations I use are also used in the dictionary with different frequency: colloquial marking ('ugs.') is used 12083 times, elevated marking ('geh.') 3905 times and specialist language marking ('fachspr.') 1487 times. In the following, I use *formal* as a cover term for elevated and specialist language. Only about 4% of the dictionary entries are marked as either colloquial or formal as the total number of entries in the dictionary is given on its back cover as *more than 500,000*.[3] This indicates that the dictionary authors use the terms only if there is a very clear intuition of a morpheme belonging to a marked register.[4]

To validate the SOLT, I planned to follow the following procedure: I randomly select 20 entries of the three categories formal, unmarked and colloquial from the dictionary. But I want to avoid entries that occur only very rarely in the two corpora I use. As a lower frequency threshold I adopt the frequency of the numeral *dreizehn* ('thirteen') with 965 occurrences in sum (689 in kern, 276 in untertitel). To randomly select formal and colloquial forms, I generate a random number n in the respective range and find the n-th occurrence of the respective mark.[5] I reject an occurrence in the following circumstances: the marking applies to only a specific interpretation of the entry, the entry is already used, the entry is less frequent than the threshold, or the entry doesn't occur at all in one of kern or untertitel. If I reject an entry, I generate a new random number. For the

[2] German also has derived forms such as *einzigste* ('only-est') that are felt to be colloquial by native speakers, though neither of the component morphemes is. Derived forms in general do not occur as entries in the dictionary, so I put aside such forms.

[3] These numbers are rather approximate especially since frequently only a specific form or meaning of a lemma is annotated for style. E.g. *Anzahl* ('number') is unmarked, but the plural *Anzahlen* is marked as specialist language.

[4] The dictionary does not describe on what basis the stylistic annotations were assigned. The dictionary states that it also heavily relies on a corpus of the Duden Verlag, which I didn't have access to.

unmarked forms I randomly generate a page number between 77 and 2115 and take the first unmarked entry fully on that page, also rejecting those marked with one of the rare stylistic markings. If the first entry has to be rejected, I move on to the next until one is accepted or the page is exhausted. In the latter case, I move on to a new random page. If unforeseen other reasons to reject a sample entry occur in the above process, I would report these in detail in the results.

After I obtain the 60 entries, I compute the SOLT of each with a python script and compile a table. Our hypothesis is that the SOLT captures to some extent native speaker knowledge of the register of dictionary entry. Therefore we expect that SOLT(formal) > SOLT(unmarked) > SOLT(colloquial). To test the prediction I apply a linear mixed model with 'entry' as a random factor and 'marking' as three-level fixed factor using the lme4 library of R and test with the function calls below whether the inclusion of 'marking' significantly improved model fit. At this point, i.e. prior to the collection of the entries, I preregistered the study (https://doi.org/10.17605/OSF.IO/H8FSV).

```
dwds <- read.csv("dwdscounts.csv", header = TRUE)
lm0 <- glmer(solt ~ (1 | lemma), data=dwds)
lm1 <- glmer(solt ~ (1 | lemma) + marking, data=dwds)
anova(lm0,lm1)
```

The following possible modification occurred to me *post hoc* as more intuitive: instead of testing a model with marking as a predictor of the SOLT, one could use the SOLT as a predictor for the marking. However, I expect both results to be driven by the correlation of the two measurements, and therefore did not explore the other direction.

1.2 Results of the Validation Study

After the start of the data collection, I observed that the frequency threshold was set to high, and lowered it to 750. Nevertheless, I could use only entries marked as 'elevated'. Of the 'specialist language' marked words as far as I could tell none was more frequent than the lower threshold, the most frequent were *Glühlampe* ('light bulb'), *kumulativ* ('cumulative'), *rezeptiv* ('receptive'), and *Kosmologie* ('cosmology'). The genitive forms of some German pronouns like *deiner* (2SG.GEN) and *unser* (1PL.GEN), are marked as elevated (and very frequent). But I did not include these for two reasons: For one, other genitive pronouns like *euer* (2PL.GEN) are not marked as elevated. And furthermore, the possessive adjectives like *dein* (2SG.POSS) and *unser* (1PL.POSS) were listed separately and not marked as elevated, but are homophonous in the Genitive to the pronouns. The data table listing all forms used in the analysis is included as an appendix.

As planned, I analysed the data statistically using R (R Core Team 2013) and lme4 (Bates et al. 2015). Specifically, I compared two models: Model 1 only included the item as a random factor to predict the SOLT value. Model 2 included in addition the Duden's stylistic marking as a fixed factor to predict SOLT. Table 1 shows that Model 2 has a better fit to the data by all criteria, e.g. the 6.5 increase in Log Likelihood represents a

[5] For 'formal', random numbers below 3906 would point to occurrences of 'elevated', those above after subtraction of 3905 to occurrences of 'specialist language'.

	Model 1	Model 2
(Intercept)	0.74*	−0.67
	(0.35)	(0.57)
marking=elevated		2.66***
		(0.80)
marking=unmarked		1.57
		(0.80)
AIC	295.91	286.92
BIC	302.19	297.39
Log Likelihood	−144.96	−138.46
Num. obs.	60	60
Num. groups: lemma	20	20
Var: number (Intercept)	0.00	0.00
Var: Residual	7.44	6.44

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Comparison of baseline statistical model with one including stylistic annotations (table generated using texreg, Leifeld 2014), the Intercept represents marking=colloquial in Model 2..

significant ($p < .005$) difference. The intercepts reported in Table 1 show also that the markings correlate with SOLT in the expected directions.

2 Quantificational Expressions

In this section, I explore how the SOLT can be applied to investigate register phenomena with quantificational expressions. My goal is specifically to use the SOLT to test the hypothesis that vagueness and/or imprecision correlate with less formal registers. I consider both cardinal numbers and other quantificational expressions. My preliminary conclusion though is that the SOLT provides no data contradicting the hypothesis, but it only provides weak evidence in favor of it.

Cardinal numbers vary by roundness (Solt 2015b and others). Solt et al. (2017) show that round numbers are cognitively easier, which might also interact with register. We explored the SOLT of the number words from 0 to 50 written in the the lowercase Roman alphabet. We did not look at higher numbers because *einundvierzig* ('forty-one') has only 37 occurrences in total, and twelve other numbers below fifty have less than 100 occurrences in the corpora. Figure 1 shows the SOLT of the cardinals from 0 to 50. The visual inspection shows a difference in between numbers up to 13 and number 14 and greater. This may reflect that numbers up to *zwölf* ('twelve') are monomorphemic in German while the number between 13 and 50 are not.[6] But the SOLT data from cardinals do not show an obvious link between roundness and register. The numbers *hundert* ('hundred') and *tausend* ('thousand') with SOLT 0.95 and 0.23 respectively provide support for this link.

[6] The bimorphemic *Dreizehn* ('thirteen') with SOLT 0.64 seems to be an outlier possibly due to its occurrence in fixed expressions such as *jetzt schlägt's dreizehn* ('I'm very surprised.'. lit. 'Now it rings 13.').

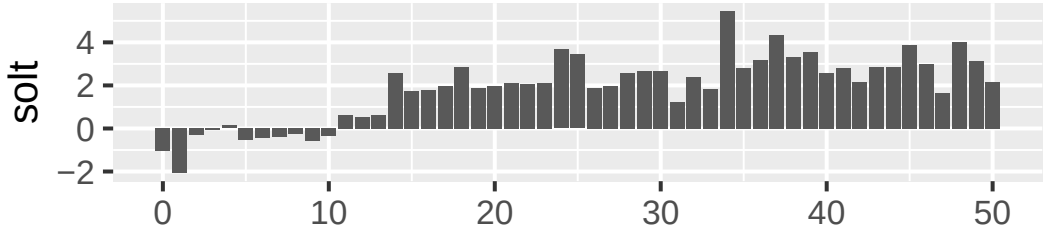

Figure 1: SOLT values for the cardinal numbers from 0 to 50.

Other quantificational expressions that are spelled as a single word constitute a small finite set which makes it difficult to test the hypothesis statistically. I furthermore exclude quantifiers that are infrequent (less than 10 000 occurrences) such as *jegliche* ('all'), *jedwede* ('all'), irgendwelche ('any-which'), *unzählige* ('uncountable') or homophonous with other forms such as *lauter* ('various many', also 'louder'). The following table shows the remaining quantificational words I looked at in four topical groups with their SOLT value .

(2) **Universals/Multitudes:** viele ('many') -0.94, alle ('all') -0.49, jede ('every') 0.06, meiste ('most') 0.24, zahlreiche ('many') 4.54
Antitones: keine ('no') -1.23, niemand ('noone') -1.65, wenige ('few') 1.20
Paucals: paar ('couple') -1.74, einige ('some (pl.)', 'unify') 0.89, manche ('some (pl.)') 1.21, mehrere ('some (pl.)') 2.06
Singulars/Definites: irgendwer ('any-who') -4.44, irgendein ('anyone') -0.83, ein ('a') -0.11, diese ('this') 0.36, jene ('that') 3.28, der ('the') 0.75

In the first two groups the (positive and negative) universals may be considered the least vague, but there is no clear support for the hypothesis that these should have higher register. The paucals are interesting because Hörmann (1983) reports experimental data that *ein paar*, *einige*, and *mehrere* are essentially synonymous. But their SOLT indicates that they differ substantially in register. This variation, however, must be independent of the vagueness hypothesis we are investigating. The singulars provide support for the hypothesis, specifically the ordering by SOLT of *ein* *irgendein* *irgendwer*. I consider the free choice forms more vague since they can mark either the speaker's ignorance or indifference.

3 Conclusion

In this paper, I explored the question of register variation of numerical and quantificational expressions from a corpus perspective. In Section 1, I proposed and validated the SOLT as a quantitative measure that correlates with register in German. In Section 2, I applied the SOLT to German quantificational expressions to test the hypothesis that vague expressions are generally associated with a less formal register. As I report, the data are mostly neutral with respect to the hypothesis. Only the higher numerals 100 and 1000 and free-choice marking on indefinites exhibit a pattern in the predicted direction. I conclude that the SOLT is a useful instrument to quantify register. But the investigation of register of

quantificational expressions specifically requires more than the tools I could apply in this paper.

References

Bates, Douglas, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai & Gabor Grothendieck. 2015. lme4: Linear mixed-effects models using 'Eigen' and S4, version 1.1-10. R Foundation for Statistical Computing. https://doi.org/10.18637/jss.v067.i01.

Brown, Gillian R. & George Yule. 1983. *Discourse analysis*. Cambridge, UK: Cambridge University Press.

Burnett, Heather. 2017. Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy* https://doi.org/10.1007/s10988-018-9254-y.

Coupland, Nikolas. 2007. *Style: Language variation and identity*. Cambridge: Cambridge University Press.

Dudenredaktion. 2015. *Duden - Deutsches Universalwörterbuch: Das umfassende Bedeutungswörterbuch der deutschen Gegenwartssprache*. Duden Verlag.

Duffley, Patrick J & Pierre Larrivée. 2012. Exploring the relation between the qualitative and quantitative uses of the determiner some. *English Language and Linguistics* 16(1). 131. https://doi.org/10.1017/S1360674311000311.

Hörmann, Hans. 1983. The calculating listener or how many are einige, mehrere, and ein paar (some, several, and a few). In *Meaning, use, and interpretation of language*, 221–234. Berlin: de Gruyter. https://doi.org/10.1515/9783110852820.

Leifeld, Philip. 2014. *texreg: Conversion of statistical model output in R to LaTeX and HTML tables*. https://doi.org/10.18637/jss.v055.i08. http://CRAN.R-project.org/package=texreg. R package version 1.33.

R Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. http://www.R-project.org/.

Smith, E, Kathleen Currie Hall & Benjamin Munson. 2010. Bringing semantics to sociophonetics: Social variables and secondary entailments. *Laboratory Phonology* 1(1). 121–155. https://doi.org/10.1515/labphon.2010.007.

Solt, Stephanie. 2006. Why *a few*? and why not **a many*. In *Proceedings of Sinn und Bedeutung 10*. 333–346. https://doi.org/10.18148/sub/2006.v10i2.736.

Solt, Stephanie. 2009. *The semantics of adjectives of quantity*. New York, N.Y.: The City University of New York dissertation.

Solt, Stephanie. 2014. An alternative theory of imprecision. In *Semantics and linguistic theory*, vol. 24. 514–533.

Solt, Stephanie. 2015a. Q-adjectives and the semantics of quantity. *Journal of Semantics* 32. 221–273. https://doi.org/10.1093/jos/fft018.

Solt, Stephanie. 2015b. Vagueness and imprecision: Empirical foundations. *Annual Review of Linguistics* 1. 107–127. https://doi.org/10.1146/annurev-linguist-030514-125150.

Solt, Stephanie, Chris Cummins & Marijan Palmović. 2017. The preference for approximation. *International Review of Pragmatics* 9. 248–268. https://doi.org/10.1163/18773109-00901010.

Solt, Stephanie & Jon Stevens. 2018. *Some three students*: towards a unified account of *some*. In *Semantics and linguistic theory*, vol. 28. 345–365. https://doi.org/10.3765/salt.v28i0.4436.

Stevens, Jon & Stephanie Solt. 2018. The semantics and pragmatics of "some 27 arrests". *University of Pennsylvania Working Papers in Linguistics* 24(1:21).

Data of the Validation Study

number	lemma	marking	kern	untertitel	total	SOLT
1	anstreben	elevated	1925	108	2033	3.472466777
2	Wiederkehr	elevated	803	75	878	2.737149035
3	behaglich	elevated	760	59	819	3.003924107
4	erachten	elevated	1291	97	1388	3.051071991
5	erschaffen	elevated	393	1852	2245	-2.919771333
6	fernhalten	elevated	733	699	1432	-0.614767709
7	Geschehnis	elevated	664	100	764	2.04789479
8	zurückkehren	elevated	4463	3121	7584	-0.167283002
9	indessen	elevated	2628	18	2646	6.506536107
10	zugehören	elevated	618	729	1347	-0.921600428
11	mithin	elevated	923	2	925	8.166898386
12	ausnehmend	elevated	1676	402	2078	1.376466291
13	obschon	elevated	749	20	769	4.543605362
14	zudem	elevated	2974	321	3295	2.528470993
15	überaus	elevated	1983	339	2322	1.865039047
16	überbringen	elevated	568	541	1109	-0.613026116
17	umher	elevated	1437	398	1835	1.168931274
18	Bemühen	elevated	1193	100	1293	2.893233686
19	verkünden	elevated	2805	569	3374	1.618211762
20	vollbringen	elevated	982	539	1521	0.1821493
1	abschießen	unmarked	1578	557	2135	0.819059521
2	Auslegung	unmarked	1401	29	1430	4.910971794
3	Bündnis	unmarked	2706	180	2886	3.226804576
4	erbärmlich	unmarked	412	901	1313	-1.81217122
5	Fachleute	unmarked	1001	43	1044	3.857673052
6	formal	unmarked	3012	57	3069	5.040327589
7	Freak	unmarked	31	930	961	-5.590179047
8	Gras	unmarked	1812	1589	3401	-0.493824621
9	herum	unmarked	7218	4851	12069	-0.109971488
10	oft	unmarked	26097	9830	35927	0.725332197
11	plus	unmarked	1021	752	1773	-0.242110152
12	probieren	unmarked	949	3274	4223	-2.469862781
13	respektieren	unmarked	1215	1751	2966	-1.210511222
14	Salon	unmarked	1263	370	1633	1.087969012
15	selbständig	unmarked	4855	105	4960	4.847721611
16	Sippe	unmarked	699	76	775	2.51793268
17	teilen	unmarked	9917	5140	15057	0.264846945
18	unbestimmt	unmarked	1342	103	1445	3.020379977
19	Verhaftung	unmarked	1416	609	2025	0.534018681
20	wegen	unmarked	24140	26823	50963	-0.835333379
1	aufbleiben	colloquial	23	112	135	-2.967081418

2	aufgedreht	colloquial	127	224	351	-1.501958687
3	eh	colloquial	687	2151	2838	-2.329913972
4	Eisenbahner	colloquial	293	37	330	2.302015037
5	fremdgehen	colloquial	16	88	104	-3.14272007
6	Halbe	colloquial	93	57	150	0.022980345
7	hinauswerfen	colloquial	235	99	334	0.563871875
8	hinwollen	colloquial	18	83	101	-2.888402882
9	Kopfweh	colloquial	40	107	147	-2.102827343
10	loskommen	colloquial	169	81	250	0.377740982
11	Oscar	colloquial	328	781	1109	-1.934915185
12	protzen	colloquial	87	26	113	1.059215326
13	Pulli	colloquial	52	277	329	-3.0965909
14	Rheuma	colloquial	81	47	128	0.1019727
15	Teufelskerl	colloquial	14	122	136	-3.806670867
16	übern	colloquial	137	75	212	0.185924941
17	verdreschen	colloquial	45	83	128	-1.566474787
18	verheddern	colloquial	69	54	123	-0.329651497
19	Wirtschaftswunder	colloquial	100	5	105	3.638639643
20	wuchten	colloquial	109	4	113	4.084895873