Detecting direction in interaction evidence

Sean Wallis, Survey of English Usage, University College London

April 2017

1. Introduction

I have previously argued (Wallis 2014) that interaction evidence is the most fruitful type of corpus linguistics evidence for grammatical research (and doubtless for many other areas of linguistics).

Frequency evidence, which we can write as $p(x)$, the probability of x occurring, concerns itself simply with the overall distribution of linguistic phenomenon x – such as whether informal written English has a higher proportion of interrogative clauses than formal written English. In order to calculate frequency evidence we must define x, i.e. decide how to identify interrogative clauses. We must also pick an appropriate baseline n for this evaluation, i.e. we need to decide whether to use words, clauses, or any other structure to identify locations where an interrogative clause may occur.

Interaction evidence is different. It is a statistical correlation between a decision that a writer or speaker makes at one part of a text, which we will label point *A*, and a decision at another part, point *B*. The idea is shown schematically in Figure 1. *A* and *B* are separate 'decision points' in a given relationship (e.g. lexical adjacency), which can be also considered as 'variables'.

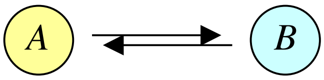

Figure 1: Associative inference from lexico-grammatical choice variable *A* to variable *B* (sketch).

This class of evidence is used in a wide range of computational algorithms. These include collocation methods, part-of-speech taggers, and probabilistic parsers. Despite the promise of interaction evidence, however, the majority of theory-driven corpus studies tend to consist of discussions of frequency differences and distributions.

In this paper I want to consider applications of interaction evidence which are made more-or-less at the same time by the same speaker/writer. In such circumstances we cannot be sure that just because *B* follows *A* in the text, the decision relating to *B* was made after the decision at *A*.

For example, in studying the premodification of noun phrases by attributive adjectives in English – which adjective is applied first in assembling an NP like *the old tall green ship*, for instance – *we cannot be sure that the adjectives are selected by the speaker in sentence order*. It is also perfectly plausible that they were selected in an alternative or parallel order in the mind of the speaker, and then assembled in the final order during the language production process.

Of course, in cases where points *A* and *B* are separated substantively in time (as in many instances of structural self-priming) or where *B* is spoken in response to *A* by another speaker (structural priming of another's language), there is unlikely to be any ambiguity about decision order. Moreover, if *A* licences *B*, then the order in unambiguous.

However, in circumstances where *A* and *B* are proximal, and where the order of decisions made by the speaker/writer cannot be presumed, we wish to consider whether there are mathematical or statistical methods for predicting the most likely order in which decisions were made.

Such a method would have considerable value in experimental design in cognitive corpus linguistics. For example, since Heads of NPs, VPs etc are conceived of as determining their complements, it may not be too much a stretch to argue that if this method is viable, we may have found a way of empirically evaluating this grammatical concept.

Howsoever desirable it may be, detecting decision-making order by *post hoc* stochastic methods is not actually possible. What this paper does is investigate methods for determining *the relative size of effect* of one variable on another and vice versa.

A and *B* may interact, but some interactions are one-sided, i.e. directional.

2. A collocation example

Let us consider a simple example 'experiment' whose result we can predict. In British English, LOOK *askance* is an archaic idiom. The adverb *askance* almost never appears without being preceded by the lemma LOOK.

However – and here is the power of an intuitive example – *the reverse is not true*. The most common words that follow LOOK are prepositions *at* (26,629) and *for* (8,117). Among adverbs, LOOK *back/forward* (at 2,170 and 2,518 respectively) or *up/down* (3,634; 2,167) are far more frequent than the rare LOOK *askance*.

The question is how we can estimate the 'one-sidedness' of the relationship between LOOK and *askance*? Using Mark Davies' interface to the *British National Corpus* (BNC), we obtain the statistics in Table 1.

	Frequency	Probability
LOOK	105,871	0.00105871
askance	48	0.00000048
LOOK *askance*	31	0.00000031
p(LOOK) × p(*askance*)		5.0818×10^{-10}
p(LOOK | *askance*) = 31/48		0.64583333
p(*askance* | LOOK) = 31/105,871		0.00029281

Table 1: Sample frequency data for LOOK, *askance*, and LOOK *askance* from the BNC[1] and some derived probabilities.

The notation 'p(LOOK | *askance*)' means the probability of the verb lemma LOOK being uttered if the following word is *askance*.

- The probability of the word *askance* being uttered in the corpus is 48 in 100,000,000 words, or 0.00000048 (or 0.000048% if you prefer). But if the previous word is LOOK, that probability is multiplied by more than 610 times, to 0.00029281.

- What about the reverse? The probability of the lemma LOOK being uttered is a little over 0.1% (0.00105871). If the following word is *askance*, that probability jumps up to nearly 65% (0.64583333). See Figure 2.

In both cases the increase is 610 times greater. If we take the ratio between the observed probability, p(LOOK *askance*) and the expected independent probability, p(LOOK) × p(*askance*), we also obtain the same ratio. So we cannot rely on the factorial increase in probability (the **odds ratio**) or any derivative of this formula (such as 'log odds' or mutual information) to give us directional information.

The reader might guess that a better method is to consider the **difference in probability**. We turn to this question in section 2.1 below.

[1] For the sake of argument we will treat the number of words as a baseline for each item (approximated to 100 million). This is not a good baseline for many purposes (Wallis forthcoming), but for the sake of our collocation example, it is fine. Strictly, we should also subtract the number of sentences for the pair LOOK *askance*, because in a sentence of length *l*, there are *l*-1 positions where LOOK can be followed by *askance*. But for example purposes we are really only interested in an order-of-magnitude calculation.

These per million word statistics are exposure statistics.

- If we hear LOOK, the probability of the next word being *askance* increases by around 0.03%. Although the probability increases, this low overall probability would not cause us to 'expect' it. LOOK *at* or LOOK *for* is far more likely (33% compared to 0.03%).
- But if we misheard the verb, and then heard the word *askance*, the chance of the previous word being *look*, *looks*, *looked*, or *looking* would increase to nearly 65%. Our brains might 'fill in' for our ears.

So far we have simply compared naïve probabilities. We have not considered whether observed probabilities are significantly different or whether the Binomial model might give us more information.

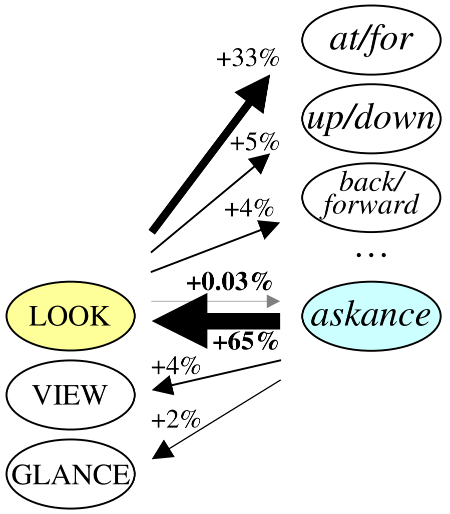

Figure 2: Sketched joint distribution of alternative words relative to the other word, for look and askance, BNC data.

2.1 Employing chi-square and phi

The first step is to employ the 2×2 χ^2 test for independence (also known as the test for 'homogeneity'). This test (Wallis 2013a) is an associative statistic, i.e. it is bi-directional. It tells us whether LOOK and *askance* appear together more frequently than would be expected by chance. Cramér's ϕ, being based on χ^2, is also associative. Table 2a obtains a 2×2 χ^2 of 18,868.63 and a (small) Cramér's ϕ of 0.01.[2]

data		independent variable		
		LOOK	**¬LOOK**	**total**
dependent variable	***askance***	**31**	17	**48**
	¬askance	105,840	99,894,112	99,999,952
	total	**105,871**	99,894,129	**100,000,000**

Table 2a: 2×2 contingency table for LOOK and *askance*.

2.2 Directional statistics

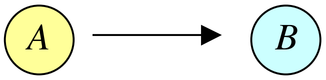

Figure 3: Directional inference from lexico-grammatical choice variable *A* to variable *B* (sketch).

However, the goodness of fit χ^2 statistic (Wallis 2013a, b) may be used in a directional way. It can be used to evaluate an increase in probability from a superset to a subset. In this case we compare the probability of the event occurring in the superset (here, the rate across the entire corpus for *askance*) with the same probability in a subset (in this example, the adverb following LOOK).

Comparing the first column in this table with the 'total' evaluates χ^2(*askance* | LOOK), whether the presence of LOOK affects the chance of *askance* appearing.[3]

We test whether p(*askance* | LOOK) differs from a given probability p(*askance*), where all the uncertainty is in p(*askance* | LOOK). This obtains a chi-square of 18,848.65 and a probabilistically-weighted goodness of fit ϕ_p (Wallis 2012) of 0.00.

[2] Input data into the 2×2 spreadsheet (tip: use the 'known totals' page) to achieve this.
[3] In the spreadsheet this 2x1 test is computed in the second row, below the 2 × 2 tests.

	LOOK	¬LOOK	total
***p*(*askance*)**	0.00029281		0.00000048

Table 2b: supplementary column probabilties for the contingency table for LOOK and *askance*.

This comparison is illustrated visually in Figure 4, which uses a 95% Wilson score confidence interval (Wallis 2013a, b). Although the probability increases 610 times from the near-zero starting point, the intervals are relatively wide and the end probability is still low, hence the tiny ϕ_p score.

data		independent variable		
		askance	**¬*askance***	**total**
dependent	**LOOK**	**31**	105,840	**105,871**
variable	**¬LOOK**	17	99,894,112	99,894,129
	total	**48**	99,999,952	**100,000,000**
	p(LOOK)	0.64583333		0.00105871

Table 2c: Transposed contingency table: now the 'independent variable' is *askance* and the 'dependent variable' LOOK.

To obtain the reverse statistic (the goodness of fit χ^2(LOOK | *askance*)), we swap rows and columns in the table (Table 2c). This obtains a χ^2 of 18,868.62 and a large ϕ_p of 0.64.

Here the χ^2 score seems only slightly elevated from the reverse statistic. However, as discussed elsewhere (see also Wallis 2013a), χ^2 scores combine size of effect and weight of evidence, so they are not good measures for comparative purposes. The key information is found in the **effect size** measure, ϕ_p. What we need to do is compare effect sizes.

2.3 Significantly directional?

We can test for **significant differences** between the two goodness of fit tests, i.e. between χ^2(*askance* | LOOK) and χ^2(LOOK | *askance*). We employ a **separability test** (Wallis 2011).[4]

Both goodness of fit tests reveal a significant difference. The separability test takes matters a stage further. It evaluates *whether the difference in these differences is significant*, i.e. whether we can report that the interaction is directional within statistical tolerances.[5]

In Table 2d, TEST 1 is the table for the goodness of fit χ^2(*askance* | LOOK); TEST 2 is the reverse test χ^2(LOOK | *askance*).

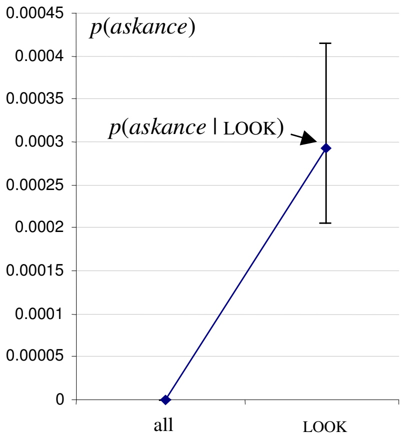

Figure 4: Increase in probability of *p*(*askance*) when the previous word is LOOK.

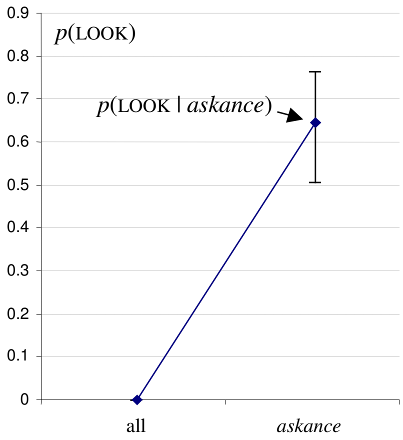

Figure 5: Increase in probability of *p*(LOOK) when the following word is *askance*.

[4] See https://corplingstats.wordpress.com/2012/03/31/comparing-c2-tests for the paper and spreadsheet.

[5] We compare simple difference *d* rather than ϕ_p. The latter is the probabilistically-weighted root mean square of differences, rather than the difference itself. Nonetheless these measures will tend to be close in a 2 × 2 table.

TEST 1			
	LOOK	¬LOOK	total
askance	31	17	48
¬*askance*	105,840	99,894,112	99,999,952
total	105,871	99,894,129	100,000,000

TEST 2			
	askance	¬*askance*	total
LOOK	31	105,840	105,871
¬ LOOK	17	99,894,112	99,894,129
total	48	99,999,952	100,000,000

Table 2d: Performing a separability test on the two goodness of fit tests. The differences d_1 and d_2 are computed between the probabilities derived from the first and 'total' column in each case. The test compares the difference in differences with an interval derived from inner Wilson intervals (Wallis 2011).[6]

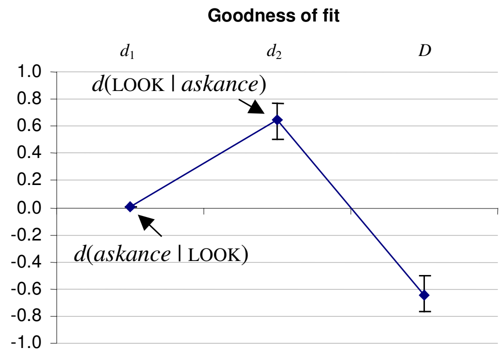

Figure 6: Comparing the difference in increases in either direction with a separability test. If D does not overlap zero the association is significantly directional.

Figure 6 plots a graph of differences with Wilson-based intervals. No intervals overlap with zero, i.e. all differences are significantly different from zero. Were we to repeat the same experiment 20 times, we would expect to see a non-zero difference in each case at least 19 times out of 20. However, difference d_1 = d(*askance* | LOOK), is tiny compared to d_2 (the increase for LOOK | *askance*).

The difference of differences, $D = d_1 - d_2$, is negative, confirming that $d_1 < d_2$. Hence we can report that both interactions are significant but also that the interaction is significantly directional.

3. A grammatical example

The example we saw in Section 2 is a lexical one with a number of features. The word-based baseline is far from ideal — no-one really believes that every word in a corpus could be *askance* — and there is a positive interaction in both directions.

Let us consider another example, taken from a forthcoming book chapter. Consider how we might evaluate the interaction of the polarity of a question tag ('TAGQ') with the polarity of the preceding verb phrase within a host clause. Compare:

David turned up did he?	VP+, TAGQ+
That's enough isn't it?	VP+, TAGQ–

Data for each pattern is obtained from the ICE-GB corpus using a single Fuzzy Tree Fragment. This obtains Table 3a. Note that the baseline in this case is all VPs followed by question tags.

data		VP		
		negative	positive	total
TAGQ	negative	2	487	489
	positive	58	172	230
	total	60	659	719

Table 3a: 2×2 contingency table for the polarity of verb phrases and question tags.

The associative χ^2 test is significant. In other words, in cases where a question tag follows a verb phrase, the polarity of the question tag is not independent from the polarity of the original VP.

[6] Data may be entered on the '2x1 goodness of fit' tab in the separability test spreadsheet.

3.1 Testing for direction under alternation

How do we test for directionality when a variable can freely vary from 0 to 1 and where both values (negative and positive) should be considered?

In our first example, we used a 2×2 χ^2 test for homogeneity (association) to compare the two variables A = {LOOK, ¬LOOK}, B = {*askance*, ¬*askance*}. The test compares both values of each variable, i.e. $\{a, \neg a\} \times \{b. \neg b\}$.

We then used goodness of fit tests to examine the changing probability of selecting a word. Our method compared $d_1 \neq d_2$ where

$$d_1 = p(b \mid a) - p(b) = p(askance \mid \text{LOOK}) - p(askance), \text{ and}$$
$$d_2 = p(a \mid b) - p(a) = p(\text{LOOK} \mid askance) - p(\text{LOOK}).$$

In this second test we only tested one value of both variables – the chance of selecting the word, $p(a)$ and $p(b)$. We did not consider the chance of selecting any other word. We did not compare, for instance, $p(\neg a)$ and $p(b)$:

$$d_1 = p(b \mid \neg a) - p(b) = p(askance \mid \neg\text{LOOK}) - p(askance), \text{ and}$$
$$d_2 = p(\neg a \mid b) - p(\neg a) = p(\neg\text{LOOK} \mid askance) - p(\neg\text{LOOK}).$$

This seems intuitive in this case: surely this doesn't matter – all values of $p(\neg word)$ except $p(\neg\text{LOOK} \mid askance)$ are likely to be close to 1! The fact that we don't even consider this prospect is probably a consequence of the fact that these values are not freely alternating.

If we employ this method in the grammatical example, however, we get four distinct results for each combination $\{a, \neg a\} \times \{b. \neg b\}$. This is summarised visually by Figure 7.

We are not weighting all cells in the contingency table equally. We obtain different results depending on which we pick. Three out of four represent a significant difference, and one is not significant. Which should we choose?

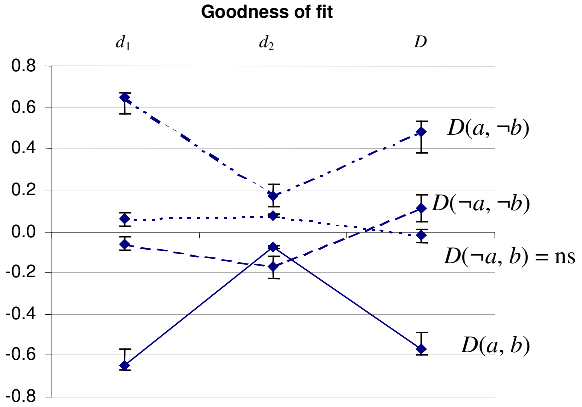

Figure 7: Different values of D obtained by comparing goodness of fit tests, where a is the freely alternating grammatical variable VP = {negative, positive} and b is the variable TAGQ = {negative, positive}.

3.2 Comparing Newcombe-Wilson intervals for direction

Instead of comparing goodness of fit tests, we propose to compare two Newcombe-Wilson tests (Wallis 2013b). This method tests if $d_1 \neq d_2$ where

$$d_1 = p(b \mid \neg a) - p(b \mid a), \text{ and}$$
$$d_2 = p(a \mid \neg b) - p(a \mid b).$$

Table 3b summarises the paired test for homogeneity.[7] Individual Newcombe-Wilson tests are significant, that is, we can report that the polarity of the question tag affects the polarity of the VP (TEST 1), and the decision of whether to employ a positive or negative VP has an effect on the polarity of the question tag (TEST 2).

[7] To achieve this import data into the spreadsheet for testing separability between two 2×2 homogeneity tests (select the '2x2 homogeneity' tab).

TEST 1	VP		
	negative	positive	total
TAGQ negative	2	487	489
positive	58	172	230
total	60	659	719

TEST 2	TAGQ		
	negative	positive	total
VP negative	2	58	60
positive	487	172	659
total	489	230	719

Table 3b: Performing a separability test on the two Newcombe-Wilson tests. The differences d_1 and d_2 are computed between the probabilities derived from the first and second column (‘negative’ and ‘positive’: not the ‘total’) in each case. The test compares the difference in differences with an interval derived from inner Newcombe-Wilson intervals (Wallis 2011).

These effect sizes are not equal. At a 5% error level, the separability test shows that the difference in these two differences is negative, i.e. the VP polarity has a significantly greater effect on the question tag (d_1 = d(TAGQ | VP) = -0.7057) than the other way around (d_2 = d(VP | TAGQ) = -0.2481). The difference in differences, D, is -0.4576, which is outside the 95% Newcombe-based interval (-0.1061, 0.0664) on the difference.

The fact that the VP precedes the question tag might lead us to expect that the interaction was directional from VP to question tag. In this case it turns out that the effect is significantly greater in this direction.

Note that in the lexical example, the direction of influence flowed in the opposite direction to word order. As we noted at the start, we cannot always take directionality for granted, especially in the case of small phrases and clauses. Our empirical method may appear to confirm our linguistic intuitions, but it is actually telling us something that we did not know until we tested it.

3.3 Optimising the difference interval

The standard separability test (Wallis 2011) compares results drawn from independent populations, such as two runs of the same experiment with data from different sources (e.g. alternative corpora), or two different experimental designs drawing data from the same source.

It combines a method evaluated by Zou and Donner (2008), which estimates the difference with the Bienaymé approximation. Essentially, as the two differences are considered to be independent, we treat them as acting at right angles in Cartesian space.

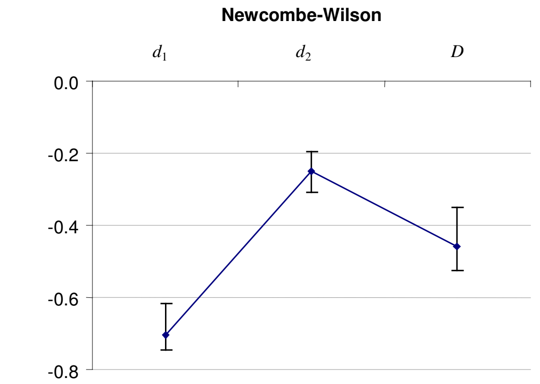

Figure 8: The significant difference, D, between two Newcombe-Wilson tests. The directional effect of the first variable, VP, is greater than the effect of the second variable, TAGQ.

$$w_D^- = -\sqrt{(w_{d_1}^-)^2 + (w_{d_2}^+)^2}\ \text{, and}$$
$$w_D^+ = \sqrt{(w_{d_1}^+)^2 + (w_{d_2}^-)^2}\ ,$$

where $w_{d_1}^-$ represents the lower Newcombe-Wilson interval width for d_1, etc.

However, in this particular application, differences and intervals are calculated from exactly the same data in the same table. They are directly coupled algebraically, but with different gradients (Figure 9): if d_1 increases, d_2 increases, and vice versa.

The Bienyamé approximation is therefore conservative. Rather, since d_1 determines d_2, we propose to take the greater of the two interval widths, i.e.

$$w_D^- = -\max(w_{d_1}^-, w_{d_2}^+), \text{ and}$$
$$w_D^+ = \max(w_{d_1}^+, w_{d_2}^-).$$

A further potential adjustment employs the continuity-corrected version of the Newcombe-Wilson interval (Wallis 2013b). This corrects for the rounding effect of the Normal approximation to the Binomial. This is generally unnecessary with the larger n where the direction test is significant.

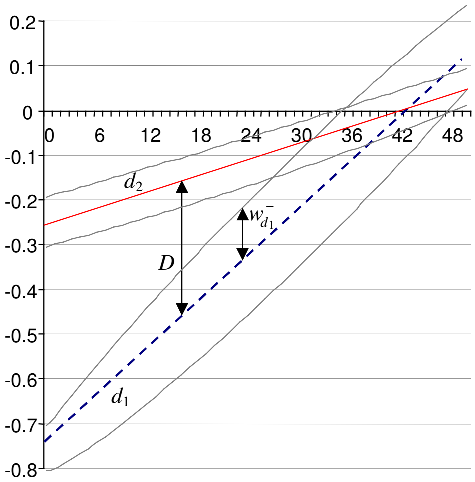

Figure 9: Relationship between d_1 and d_2, with NW intervals, as the top left cell in Table 3a permutes from 0 to 50, transferring data to the cell below.

4. Mapping significance of association and direction

Figure 10 shows the relationship between the associative χ^2 test and the direction test (separability of Newcombe-Wilson differences). Using the method described in (Wallis 2013b) we compute all integer combinations of a 2 × 2 table where both rows n_1 and n_2 sum to 400. We test for association and significant difference, and mark the grid accordingly in Figure 10. We visualise the effect of employing both the Bienyamé and 'max' interval estimate.

Our first observation is that (perhaps unsurprisingly) we need more data to determine direction than to detect simple association. See Figure 11. With a table summing to $n = 20$ in both rows, there is no observed combination that could be said to imply direction. We have insufficient data.

With both rows summing to 100, only 9.65% of combinations[8] are directional using the max formula (6.59% for Bienaymé). This proportion increases to 28.82% (23.43%) when n_1 and n_2 increase to 500, and 45.16% (39.34%) at 2,000.

The more data employed, the smaller the confidence intervals will be, and the greater the likelihood that differences in differences will be significant.

The four areas represent positive and negative significant differences in both directions, i.e. the combinations $\{a, \neg a\} \times \{b. \neg b\}$.

The shape and location of these areas is also interesting. The greatest directional difference is not where we see the greatest skew (Cramér's ϕ, Wallis 2013a) but when one cell is close to zero and an adjacent cell is small.

In a maximally-skewed contingency table, [[0 n] [n 0]] (Figure 10, top left or bottom right), both variables impact on each other – maximally *and*

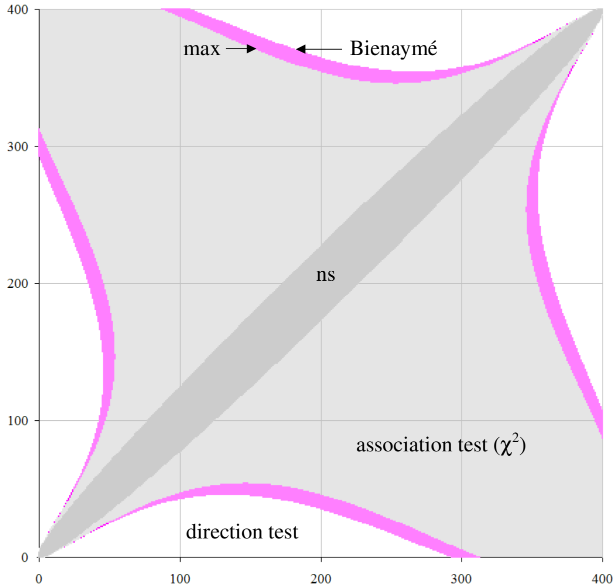

Figure 10: Association and direction test outcomes for 401^2 possible combinations ($\alpha = 0.05$, $n_1 = n_2 = 400$).

[8] This is a proportion of *combinations*, but not all combinations are equally likely. The total prior Binomial probability of these areas (for P=0.5) is infinitesimal. By comparison, the area marked 'ns' in Figure 9 represents 95% of outcomes.

equally – with the result that there is no significant difference in direction. As a result, the direction test contour looks very different than that for association.

5. Concluding remarks

This evaluation performs three different significance tests, one of which is employed twice.

- The 2×2 χ^2 test simply tests for **association**, i.e. whether the two variables (outcomes of choices at *A* and *B*) interact.
- The second test (the 2×1 goodness of fit χ^2 test or the Newcombe-Wilson test) obtains **directional** information. It is used twice:
  - to test whether making a particular choice at *A* correlates with the chance of making a particular choice at *B* to significantly increase; and
  - to test whether making a choice at point *B* correlates with an increased propensity to make a particular choice at point *A*.
- These two tests are then contrasted with a **separability test**. This evaluates whether the increase in one direction is significantly greater than in the other.

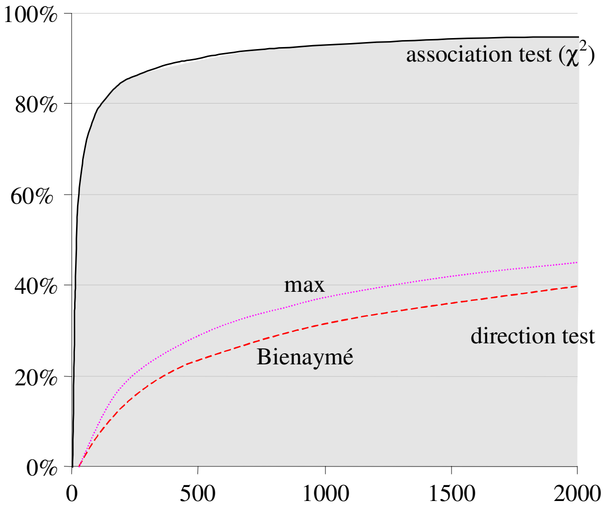

Figure 11: Percentage of tables with $(n+1)^2$ combinations that show a significant association, and those where effect sizes can be said to be significantly different ($\alpha = 0.05$, $n = n_1 = n_2$).

The first example we used above was based on a simple lexical collocation, and we used the goodness of fit test on the basis that the two outcomes – to select a word and not to select a word – are conceptually unequal. The goodness of fit method focuses on a particular cell in the table (here, where both LOOK and *askance* are uttered together). However, this method is not appropriate where variables are free to vary, probabilities can range from 0 to 1, and we are interested in how the variable behaves overall.

The principles outlined here are extensible to the detection of interaction directionality between decisions in any process of language production, as the second example should make clear. In order to address the limitations of the goodness of fit method, we compare the difference in selection probability for each variable using the Newcombe-Wilson method.

In Wallis (2013b) we remarked that the Newcombe-Wilson test was preferable to the χ^2 test to compare probabilities drawn from independent populations, and indicated that this did not include grammatical interaction (as speakers could in principle choose to express both items).

This is because this class of test compares variation computed independently for each value of *p* (and thus, in our case, for each variable) rather than jointly across the 2 × 2 grid. It derives from the paired Binomial test rather than the Fisher test. At the limit of significance the outcome converges, i.e. the tests almost always obtain the same result. In the current application we use the confidence intervals of this test in a different way: to compare sizes of effect in each direction.

These assessments of difference (d_1 and d_2) can obtain significantly different results, but only if the associative test is not borderline significant.[9] If we do obtain a significant difference in differences, we can say that one variable appears to have a greater effect on the other, i.e. we have evidence of

[9] At the very edge of significance, i.e. where the associative 2×2 χ^2 test is just significant, both Newcombe-Wilson tests will obtain a very similar result (try overtyping the cell in the grammatical example marked '2' with '114') – so the difference in differences will be non-significant.

directional influence.

Throughout this paper, I avoided claiming that we are detecting **causation**. We are finding evidence that we might expect to see in patterns of one-way causation, but this does not mean that the underlying reason for this pattern is one of cause-and-effect. Statistics can only detect correlations, not causes.

A correlation between two variables may derive from a third variable affecting both variables in a particular way. To borrow an example from Stephen J. Gould, petrol prices may rise alongside the age of a petrol pump attendant, but we would not claim that one causes the other!

It is particularly important to carefully consider directional results for similar reasons. In this paradigm, results satisfying the direction test are those where two variables correlate, but we can also say specifically, *that one changes by a significantly greater extent than the other*. As with all correlations, this is a distributional observation. It means that an observed greater size of effect in a particular direction is unlikely to have occurred by chance.

Conceptually, this issue also relates to the question of **freedom of variation** (i.e. that all cases of A could, in theory at least, be a or $\neg a$, so $p(a)$ can range from 0 to 1). If both decisions are free to vary, then both differences are on the same probabilistic scale and a comparison of differences by subtraction is a legitimate operation.

A refinement to our lexical example (section 2), would therefore consist of grammatically restricting data to (for example) verb+adverb sequences, and obtain baselines and probabilities accordingly. This exercise is left to the reader to undertake.

References

Wallis, S.A. 2011. *Comparing χ^2 tests for separability*. London: Survey of English Usage, UCL.

Wallis, S.A. 2012. *Goodness of fit measures for discrete categorical data*. London: Survey of English Usage, UCL.

Wallis, S.A. 2013a. z-squared: the origin and application of χ^2. *Journal of Quantitative Linguistics* **20**:4, 350-378.

Wallis, S.A. 2013b. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics* **20**:3, 178-208.

Wallis, S.A. 2014. What might a corpus of parsed spoken data tell us about language? In L. Veselovská and M. Janebová (eds.) *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*. Olomouc: Palacký University, 2014. pp 641-662.

Wallis, S.A. forthcoming. That vexed problem of choice. London: Survey of English Usage, UCL.

Zou G.Y. and Donner A. 2008. Construction of confidence limits about effect measures: A general approach. *Statistics in Medicine* **27**: 1693-1702.