Statistik/Signifikanz: Die Projektleitung hat sich für eine Aufweichung von unseren Regeln bzgl. des strikten Befolgens der Signifikanz vor allem aber der erwarteten Werte ausgesprochen, weil unsere Methode, wie sich immer wieder gezeigt hat, nicht für alle Fälle geeignet ist (bspw. für reine Helvetismen), es aber ebenso wenig sinnvoll bzw. möglich ist, für jeden Fall die passende Methode heranzuziehen. **Ich (Gerard Adarve) liste in der Folge zur Rekonstruktion dieser Diskussion und zu Ihrer/Eurer Information die folgenden vier Mail-Sequenzen weiter: a) Mail von Don Tuggener vom 12.06., b) Mail von mir vom 13.06., c) erstes Mail der Projekleitung (Christa Dürscheid) zu diesem Thema vom 06.07., d) zweites (bisher noch nicht veröffentlichtes) Mail der Projektleitung (Stephan Elspaß) zu diesem Thema mit Präzisierungen vom 07.07.:**

a) Liebe alle,

ich melde mich hier nochmals aus statistischer Perspektive. Die Statistik soll meiner Meinung nach nicht dazu dienen, dutzende bestehende Artikel zu löschen oder Mitarbeitende davon abzuhalten, Artikel zu Phänomenen zu schreiben, die für sie (Aufgrund der Belegverteilung) offensichtlich Varianten sind. Folgende Punkte sind aus meiner Sicht relevant:

- Die **Statistik** soll generell ein **Hilfsmittel** sein beim Entscheiden, ob ein Artikel verfasst werden soll. Aber eine **gesamtheitliche Auswertung** der statistischen Ausgabe (absolute Belegzahlen, erwartete Werte, Pearson Residuals, p-Wert) in Kombination mit den **Schwellenwerten** ist unumgänglich. **Es genügt nicht, sich auf einen einzelnen Grenzwert oder Parameter festzulegen**. Daher braucht es das angesprochene 'Bauchgefühl', resp. die Expertise der/des Bearbeitenden eben. Im Übrigen sind **statistische Tests** in der **Korpuslinguistik** (mittlerweile teilweise höchst) **umstritten** - bei Bedarf gerne mehr dazu.
- Ein **statistischer Test** ist **nicht in allen Fällen notwendig und sinnvoll** (Wenn alle Belege in der Schweiz sind, was soll getestet werden? Natürlich sind dann viele der erwarteten Werte unter 5 - soll deshalb kein Artikel geschrieben werden?).
- Statistische Tests wurden entwickelt, um die **Auswertung von Grenzfällen** (kleine Unterschiede in den Daten, sind sie wirklich wichtig?) mit einem theoretischen/wissenschaftlichen Unterbau zu versehen. Die 5%- oder 1%-Schwelle im Zusammenhang mit dem p-Wert (> 0.05 für Signifikanz) wurde in der Wissenschaft auf Community-Ebene (mehr oder weniger) arbiträr aufgrund von Erfahrungswerten festgelegt (es gibt keinen mathematischen und klaren Hintergrund). Gleiches gilt für die 80-20%-Regel bei den erwarteten Werten.
- Es wäre **theoretisch problemlos zulässig**, bei einem Test **Regionen mit wenig Belegen auszuschliessen** und den Test nur über den Regionen zu machen, die für ein Phänomen relevant sind (also genug Belege haben). D.h. **viele der erwarteten Werte unter 5 sind schlicht nicht relevant**. Unsere Plattform ist leider nicht flexibel genug, solche interaktive Funktionalität (z.B. einzelne Regionen aus Tabellen löschen) zu implementieren. Das liegt zeit- und ressourcentechnisch leider nicht drin. Ich kann aber anbieten, in Einzelfällen solche Auswertungen händisch zu machen.
- Der **Chiquadrattest** ist **nicht** die **optimalste Lösung** aus korpuslinguistischer/statistischer Sicht für (alle) unsere Probleme. Idealerweise würden wir bei jedem Phänomen separat überlegen, welcher der Dutzenden von Tests am besten geeignet ist gegeben der Beleglage. Das ist nicht praktikabel. Der Chiquadrattest ist **aber der beste "Allrounder" für uns**, daher verwenden wir ihn. (Ein besserer Test für Phänomene mit wenig Belegen wäre z.B. der Fisher-Test. Der lässt sich aber nicht einfach als Webservice laufen lassen, da man ihn manuell und interaktiv konfigurieren muss.)
- Die **einzige Konstellation von Werten** in der statistischen Analyse, **die zwingend(!) zur Löschung oder zum Nicht-Schreibens eines Artikel führen muss** ist:
-- Es gibt in allen/den meisten Regionen genügend Belege, d.h. die erwarteten Werte sind (tendenziell alle) über/nahe bei 5
-- Der p-Wert ist deutlich über 0.05 - Hier kann man ohne weiter zu überlegen abbrechen. Das ist aber das einzige "harte" Kriterium, das die Statistik in unserem Setting zu bieten hat.

Als **konkretes Vorgehen im Umgang mit der Statistik** würde ich also folgende **Sanity Checks vorschlagen**:
1. Beim Betrachten der Verteilung der absoluten Werte (nach Überprüfen der festgelegten Schwellenwerte) ergibt sich, dass es keine klaren oder nur schwache Präferenzen gibt, die aber intuitiv einen Sinn (hier ist das Bauchgefühl) ergeben (Grenzfälle eben)
2. Die erwarteten Werte sind tendenziell über oder knapp bei 5
3. Die Pearson Residuals zeigen, dass eine oder mehrere Regionen Werte über (absolut, also

ungeachtet des Vorzeichens) 2 oder 3 haben.

Wenn diese drei Kriterien nicht erfüllt sind, macht es wenig Sinn, die Statistik bei der Entscheidung über Artikel/Nicht-Artikel zu berücksichtigen. Das soll aber nicht heissen, dass automatisch kein Artikel geschrieben werden darf! Es heisst nur: Die Statistik kann hierzu nichts Eindeutiges sagen.

Unser Motto sollte also nicht sein: "Alles, was einen p-Wert grösser als 0.05 / erwartete Werte unter 5 hat, kommt nicht in die VG", sondern eher: "Wenn in einer Belegverteilung gewisse schwache Tendenzen sichtbar aber unsicher sind und die Ausgangslage es erlaubt, machen wir einen Signifikanztest, bevor wir einen Artikel schreiben."

Das Zielpublikum der VG ist breit und die VG dient einem anderen Zweck, als nur den strengen (und teils umstrittenen) Leitplanken der Statistik zu folgen. Das ist beim Verfassen wissenschaftlicher Artikel anders - die Artikel in der VG sind aber nicht Beiträge in Fachzeitschriften zur statistischen Korpuslinguistik.

Soweit meine Meinung in der Sache. Ich hoffe ich konnte hier einige Punkte beleuchten und ein einigermassen konkretes Vorgehen vorschlagen, das praktikabel ist.. Ganz vom Bauchgefühl (das hoffentlich wenig Bauchweh verursacht) kommen wir nicht weg.

Ich schliesse mich Geris Schlusswort an, weil das Ganze eine methodische Frage ist:
"Falls eine solche Lösung zu "schwammig", zu undeutlich, zu missverständlich, zu wenig empirisch, zu frei, zu beliebig usw. sein sollte, dann ist das m.E. nicht ein Punkt für eine Teamdiskussion, sondern Chefsache bzw. würde ich dann die Projektleitung bitten, diesen Aspekt zu diskutieren und danach klar zu kommunizieren."

b) Sehr geehrter Herr Ziegler, lieber Stephan, liebe Christa

Die Auflockerung der statistischen Regeln insbesondere hinsichtlich der erwarteten Werte hat in den letzten Wochen einiges zu Reden gegeben und immer wieder zu Unsicherheiten und Rückfragen geführt. Ich persönlich war mit Blick auf die Einheitlichkeit bis vor dem Workshop stets Vertreter einer strikten Befolgung der statistischen Ergebnisinterpretation, komme aber, nicht zuletzt nach Gesprächen mit Don Tuggener und Mails wie seine gestrige (siehe unten; generell: alle relevanten Punkte zu dieser Diskussion können den verschiedenen Mailwechseln unten entnommen werden, auch wenn es sich nur um einen Teil davon handelt), immer mehr zum Schluss, dass es in diesem Feld, gerade bei einem Setting, wie wir es haben, keine absoluten statistischen Lösungen gibt und dass es nicht in allen Fällen und durchgehend sinnvoll ist, "blind" nach statistischen Vorgaben zu funktionieren. Es muss m.E. einen bestimmten Spielraum geben, der es den Mitarbeitenden erlaubt, sich darüber hinwegzusetzen und in manchen Fällen die persönliche Expertise bei der Interpration von Werten höher zu gewichten als die Statistik. Würden wir das nicht so handhaben, so hätte das in manchen Bereichen die Löschung eines Grossteils der Artikel zur Folge (welche in den meisten Fällen bei "gesundem Menschenverstand" eigentlich in die VG gehören würden, von der Statistik her aber schon rein strukturell gar nicht die Chance bekommen (können), alle Vorgaben – insbesondere bei den erwarteten Werten, weniger bei der Signifikanz – zu erfüllen). Dies heisst nicht gleichzeitig, dass jede/jeder nun einen Freipass bekommen sollte, um "alles" in die VG reinzunehmen, sondern dass wir die individuelle Interpretation der Einzelnen aufwerten (und ihr vertrauen) und den beschriebenen Spielraum, sich in manchen, in der Datenbank dokumentierten Fällen über die Statistik hinwegsetzen, öffnen. Entscheidend soll in solchen Fällen, wie unten immer wieder geschrieben, der "gesunde Menschenverstand", die Expertise und das Bauchgefühl der Bearbeiterin/des Bearbeiters sein.

Dieses Vorgehen ist zweifellos "schwammig", weil nicht absolut (insofern auch individuell verschieden auslegbar). Ich selber sehe da aber ehrlich gesagt keine andere Lösung, wenn wir der Materie und unserem Setting gerecht werden wollen. Statistik oder besser gesagt: unsere statistische Methode ist, wie Don Tuggener schreibt, "der beste Allrounder für uns", hat aber eben auch Schwächen, und genau da haben wir die Möglichkeit, korrigierend einzugreifen.

c) Liebe alle,

von Seiten der Projektleitung ist noch die Frage offen, ob wir nicht unsere eigenen Beschlüsse unterlaufen, wenn wir auch nicht statistisch signifikante Varianten im Wiki erfassen. Dazu möchten wir Folgendes festhalten: Wenn sich bei der Bearbeitung solchen Varianten ergibt, dass gute Gründe dafür sprechen, sie aufzunehmen, dann sollten wir das auch tun. Wir folgen hier der Meinung von Don Tuggener, der ja geschrieben hatte, dass die Statistik ein Hilfsmittel beim Entscheiden, ob ein Artikel verfasst werden soll, sein kann, dass aber eine gesamtheitliche Auswertung der statistischen Ausgabe (absolute Belegzahlen, erwartete Werte, Pearson Residuals, p-Wert) in Kombination mit den Schwellenwerten unumgänglich ist.

d) Die statistische Signifikanz des Auftretens einer Variante nach unserem Verfahren soll und kann nicht unbedingte Voraussetzung für die Aufnahme eines entsprechendes Artikels im Wiki sein. Wenn sich bei der Bearbeitung solchen Varianten ergibt, dass die areale Distribution der Variante und ihrer Gegenvariante(n) in absoluten Zahlen dafür sprechen, sie aufzunehmen, dann sollten wir das auch tun. Die Begründung dafür ist: Wenn eine bestimmte areale Distribution offensichtlich ist, aber nur aufgrund der (niedrigen) Beleglage in manchen unserern Regionen und nur anhand unseres Chi2-Tests nicht als signifikant bewertet werden kann, heißt das nicht, dass eine tatsächliche Signifikanz auszuschließen ist.

Diese Regelung erspart uns, einen Großteil der Artikel z.B. im Bereich der Genuszuordnung löschen zu müssen - obwohl etwa ganz klar ist, dass *das Tram* ein Helvetismus ist. (Don Tuggener schrieb dazu: "Wenn alle Belege in der Schweiz sind, was soll getestet werden?") Gleichwohl sollten wir diese Fälle markieren. Für den/die Normalleser/in genügt die Information, dass nach unseren Ergebnissen (auch wenn sie nicht immer den Signifikanz-Test 'bestehen') eine bestimmte Variante als areal typische Variante aufzufassen ist; wenn auf einer Karte viele Torten blaß bleiben, ist auch auf für ihn/sie deutlich, dass z.T. niedrige Belegzahlen vorliegen. Durch eine Markierung und den Verweis (mit Link) auf die Seite http://dssrv014.uzh.ch/w/index.php/Korpus,_Regionen_und_analysierte_Zeitungen#Verarbeitung_von_Varianten_und_Statistik wird für näher Interessierte offengelegt, dass das Ergebnis nur nach dem dort beschriebenen Verfahren statistisch nicht signifikant ist (aber es z.B. nach dem Fisher-Test evtl. wäre). So machen wir unser Verfahren transparent. Auch die Angabe, dass wir nach dem Chi2-Test vorgehen, ist ja ein Teil dieser Transparenz. Wie Don Tuggener auch deutlich gemacht hat, fiel unsere Wahl auf diesen Test, weil er "der beste Allrounder für uns" sei.

Abschliessender Kommentar Gerard Adarve: **Ich hoffe, dass dieses Vorgehen und der Grundgedanke, der hinter diesem Vorgehen steht, allen klar ist. Es handelt sich dabei, wie oben geschrieben, nun aber keinesfalls um einen „Freipass“, um "alles" in die VG reinzunehmen, sondern es bedeutet, „dass wir die individuelle Interpretation der Einzelnen aufwerten (und ihr vertrauen) und den beschriebenen Spielraum, sich in manchen, in der Datenbank dokumentierten Fällen über die Statistik hinwegzusetzen, öffnen. Entscheidend soll in solchen Fällen der "gesunde Menschenverstand", die Expertise und das Bauchgefühl der Bearbeiterin/des Bearbeiters sein.“ Wie Stephan Elspaß vorschlägt, werden wir uns eine eigene Markierung und eine entsprechende Passage in http://dssrv014.uzh.ch/w/index.php/Korpus,_Regionen_und_analysierte_Zeitungen#Verarbeitung_von_Varianten_und_Statistik für solche Fälle überlegen, bis dorthin wird gebeten, solche Fälle nicht nur in der Datenbank zu dokumentieren, sondern separat auch bspw. in einem Worddokument zu listen. Dies wird es uns zu einem späteren Zeitpunkt erleichtern, alle betroffenen Fälle schnell zu finden und zu markieren. Für Fragen zu diesem Thema (wie auch zu allen anderen Themen) stehe ich Ihnen/Euch selbstverständlich jederzeit zur Verfügung.**