

## FORSCHUNGSERGEBNISSE ZUR FRAGE: HABEN 80 % ALLER INFORMATIONEN EINEN RAUMBEZUG?

Stefan Hahmann, Dirk Burghardt

**Zusammenfassung:** Die These, dass 80 % aller Informationen einen Raumbezug haben, ist in der GIS-Welt weit verbreitet. Sie entstand während des Aufkommens der ersten Geoinformationssysteme und dient seitdem im gesamten Spektrum der Branche als Marketinginstrument. Jedoch findet sich bei der Recherche nach dem Ursprung der These keine Studie, die die Zahl empirisch belegt. Dieser Artikel soll dazu beitragen, diese Forschungslücke zu schließen. Hierfür stellen wir zwei Ansätze vor: einen netzwerk-basierten und einen kognitiven. Im netzwerk-basierten Ansatz nutzen wir die deutsche Wikipedia, deren Artikel wir als Knoten und deren interne Verlinkungen wir als Kanten eines gerichteten Graphen modellieren. Für den kognitiven Ansatz wurde ein Experiment durchgeführt, in dem Probanden die Aufgabe hatten, Informationen in die Kategorien „Direkter Raumbezug“, „Indirekter Raumbezug“ und „kein Raumbezug“ einzuordnen. Im Beitrag werden die Ergebnisse beider Ansätze vorgestellt. Eine anschließende Synthese liefert eine empirisch begründete Zahl für die „Raumbezugsthese“.

**Schlüsselwörter:** Raumbezug, Netzwerktheorie, räumliche Wahrnehmung, Wikipedia, Geographische Informationsgewinnung

## RESEARCH RESULTS ON THE QUESTION: IS 80% OF ALL INFORMATION GEOSPATIALLY REFERENCED?

**Abstract:** The assertion that 80% of all information is somehow geospatially referenced is widely known in the GIS community. It originated in Canada/USA during the rise of geo information systems and has become a powerful marketing tool for the whole geoinformation sector. However, there is no scientific study that empirically proves this assertion. This paper is a contribution with the aim to fill this research gap. For this purpose we introduce a network and a cognitive approach. For the network approach we use the German Wikipedia. Its articles are interpreted as nodes and its links are interpreted as edges of a directed graph. For the cognitive approach an experiment was carried out. The task of the participants was to assign information to categories of geospatial reference („direct geospatial reference“, „indirect geospatial reference“, „no geospatial reference“). This paper presents the results of both approaches. A subsequent synthesis provides an empirical justified figure for the „geospatial reference assertion“.

**Keywords:** Geospatial reference, networks theory, spatial cognition, Wikipedia, geographic information retrieval

### Autoren

Dipl.-Ing. (FH) Stefan Hahmann  
Prof. Dr.-Ing. Dirk Burghardt  
Technische Universität Dresden  
Institut für Kartographie  
Helmholtzstraße 10  
01069 Dresden

E: stefan.hahmann@tu-dresden.de  
dirk.burghardt@tu-dresden.de

## 1 EINLEITUNG

Der genaue Ursprung der These, dass 80 % aller Informationen einen Raumbezug haben, lässt sich heute nicht mehr zweifelsfrei identifizieren. In Hahmann et al. (2011) wurde über die ältesten auffindbaren Quellen berichtet. Es lässt sich feststellen, dass der Ausgangspunkt der Zahl in den 1980er-Jahren liegt, und dass die These aus Kanada/USA stammt. Eine empirische Studie, die die Zahl belegt, existiert jedoch nicht. Die frühesten Quellen zeigen, dass die These ursprünglich auf kommunale Datenbestände bezogen war, da diese viele Referenzen zu Adressdaten, Koordinaten und räumlichen Identifikationsnummern (z. B. Flurstücksnummern, Postleitzahlen) enthalten. Inzwischen wurde die Annahme jedoch (über)generalisiert und findet sich auch in allgemeinen Kontexten, nicht selten auch in wissenschaftlichen Publikationen, wie „Entsprechend der allgemein akzeptierten Grundannahme, dass 80 Prozent aller Informationen einen Raumbezug aufweisen [...]“ in Fitzke & Greve (2010, S. 735). Die wissenschaftliche Fragestellung, die sich daraus ableitet ist demzufolge: Weisen tatsächlich 80 Prozent aller Informationen einen Raumbezug auf?

Broder et al. (2000) haben für das World Wide Web gezeigt, dass – abgesehen von wenigen vollständig isolierten Webseiten – alle Seiten mit allen anderen (mittelbar) verlinkt sind. Ausgehend von dieser Erkenntnis könnte man zu der Annahme gelangen, dass für Informationen, die in einer Netzwerkstruktur gespeichert sind, 100 % die richtige Zahl sein müsste, sofern im betrachteten Netzwerk zumindest einige georäumliche Informationen gespeichert sind. Der Pfad, der innerhalb eines Informationsnetzwerks eine beliebige Information mit einer georäumlichen Information verbindet, mag dabei als Realisierung eines Georaumbezugs betrachtet werden. Wichtig ist jedoch letztendlich, für welchen Anteil an Informationen die georäumliche Komponente relevant ist. In Analogie zum Bild des Informationsnetzwerks stellt sich deshalb die Frage, wo die menschliche Wahrnehmung den Pfad zwischen einer beliebigen Information und einer georäumlichen Information trennt.

Für die Untersuchungen, die wir durchgeführt haben, wurde die deutsche Wikipedia als Referenzkorpus genutzt. Wir folgen dabei dem Paradigma von Hecht

& Moxley (2009), die Wikipedia als Sammlung fachgebietsneutralen Wissens betrachtet haben und diese bereits nutzten, um Toblers erstes Gesetz der Geographie zu überprüfen.

## 2 NETZWERK-BASIERTER ANSATZ FÜR DIE BESTIMMUNG RAUMBEZOGENER INFORMATIONEN

Der netzwerk-basierte Ansatz für die Charakterisierung von Informationen hinsichtlich ihres raumbezogenen Anteils wurde in Hahmann et al. (2011) vorgestellt und ist für die aktuellen Arbeiten zu diesem Artikel erstmals implementiert worden. In Analogie zu sozialen Netzwerken, in denen man von Kontakten ersten bis n-ten Grads spricht, wird das Konzept des Raumbezugs n-ten Grads in Informationsnetzwerken genutzt. Im Weiteren wird hierfür auch der Begriff *Netzwerkgrad des Georaumbezugs (NGGR)* verwendet. Die deutsche Wikipedia als Untersuchungsgegenstand ist geeignet, um als Netzwerk modelliert zu werden. Als primär georäumliche Infor-

koordinaten gekennzeichnet sind, werden im Weiteren als *georäumliche Artikel* bezeichnet.

Im Netzwerk der Wikipedia-Artikel stellt jeder Artikel einen Knoten dar und jeder Link zwischen zwei Artikeln eine gerichtete Kante. Das Netzwerk wird somit als *gerichteter Graph* modelliert. Für unsere Untersuchungen definieren wir, dass ein Raumbezug ersten Grads entsteht, wenn ein beliebiger Artikel mindestens einen Link zu einem georäumlichen Artikel hat. Jeder georäumliche Artikel selbst wird als Artikel mit Georaumbezug 0-ten Grads bezeichnet. Für alle weiteren Grade des Georaumbezugs gilt, dass sich der Grad des Georaumbezugs mit jedem Link, der zusätzlich benötigt wird, um innerhalb des Netzwerkes von einem beliebigem Artikel auf dem kürzesten Weg zu einem georäumlichen Artikel zu gelangen, um eins erhöht.

Für die Kanten werden keine Gewichte, beispielsweise auf Basis der Anzahl gleicher Verlinkungen, genutzt. Der betrachtete Graph kann somit weiterhin als *ungewichteter*

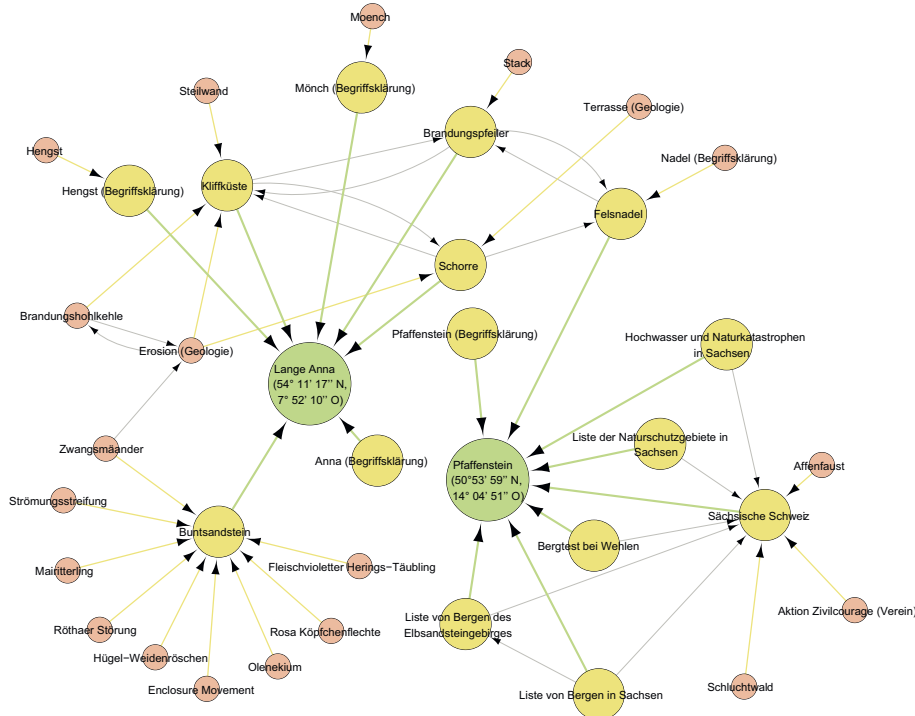


**Abbildung 1:** Artikel der deutschen Wikipedia, der eine Koordinatenvorlage in der Überschrift verwendet. Für eine ausführlichere Beschreibung des Aufbaus von Wikipedia-Artikeln mit Fokus auf Koordinaten und Verlinkungen siehe Dahinden (2011a).

mationen innerhalb dieses Graphen betrachten wir dabei ausschließlich Artikel mit Koordinaten, die durch dafür vorgesehene Vorlagen in den Überschriften der entsprechenden Artikel eingefügt wurden (siehe Abbildung 1). Dahinden (2011a) bezeichnet diese Koordinaten als „explizite Koordinaten“. Artikel, die mit solchen expliziten

koordinaten gekennzeichnet sind, werden im Weiteren als *georäumliche Artikel* bezeichnet.

Um die Koordinaten aus georäumlichen Artikeln nutzen zu können, müssen diese zunächst extrahiert werden. Dieser Schritt wird von einigen Wikipediabenut-



**Abbildung 2:** Artikel (Kreise) mit Georaumbezug nullten (grün), ersten (gelb) und zweiten (rot) Grads. Grüne Pfeile zeigen Verlinkungen an, durch die NGGR 1 entsteht, gelbe Pfeile zeigen Verlinkungen an, durch die NGGR 2 entsteht. Graue Pfeile stehen für weitere Verlinkungen zwischen Artikeln, die für den Netzwerkansatz jedoch nicht betrachtet werden.

zern automatisiert durchgeführt. Für die Untersuchungen zu diesem Beitrag eignen sich aufgrund ihrer Vollständigkeit die Koordinatendateien des Wikipediabenutzers Dispenser (2011) am besten. Diese werden täglich für jede Sprachversion der Wikipedia aktualisiert. Einige Koordinaten, die sich nicht auf die Erde bezogen, wie solche, die auf verschiedenen Artikeln über Mare und Krater des Mondes eingetragen sind, wurden gefiltert. Weiterhin zeigte sich, dass die Koordinatendateien nicht ausschließlich explizite Koordinaten, sondern auch von Dahinden (2011a) als „implizite Koordinaten“ bezeichnete Werte enthalten. Implizite Koordinaten werden beispielsweise für die Annotation von Bildern mittels Geotags genutzt oder für zusätzliche Ortsangaben im Volltext von Artikeln. Diese Koordinaten wurden ebenfalls gefiltert. Dieser Schritt ist für den Netzwerkansatz von großer Bedeutung, da die Angabe einer expliziten Koordinate für einen Artikel eine bewusst getroffene Entscheidung der Wikipedia-Autoren darstellt, dass sich die Koordinaten auf das gesamte Konzept des Artikels beziehen und somit den gesamten Artikel zu einem (geo)räumlichen Artikel machen. Da das Einfügen einer expliziten Koordinate einen kognitiven

Prozess widerspiegelt, besitzt der Netzwerkansatz also sowohl analytische als auch kognitive Anteile.

Für den Netzwerkansatz stellt sich die Frage, ob eine ungefähre Vorhersage des zu erwartenden Wertebereichs des NGGR möglich ist. In diesem Zusammenhang sind Forschungsarbeiten zu den Netzwerkstrukturen der verschiedenen Wikipedias von Interesse (z. B. Zlatić et al. 2006, Capocci et al. 2006, Voss 2005). Es konnte bereits gezeigt werden, dass sich das Netzwerk der Wikipedia-Artikel einer Sprachversion mit dem Small-World-Modell nach Watts-Strogatz (1998) beschreiben lässt. Zlatić et al. (2006) bestimmten die durchschnittliche Entfernung zwischen zwei beliebigen Artikeln empirisch. Sie fanden, dass diese im Mittel 4,53 beträgt und dass dieser Wert bei verschiedenen Sprachversionen nur eine geringe Streuung zeigt. Das Ergebnis einer eigenen Berechnung – 4,78 – mit dem für die Untersuchung verwendeten Stand der deutschen Wikipedia von 2011 bestätigt diesen Wert.

Es ist deshalb zu erwarten, dass der durchschnittliche NGGR wesentlich kleiner ist als 4,78, da für den Netzwerkansatz zur Bestimmung des Georaumbezugs nicht alle beliebigen Entfernungen zwischen

zwei Artikeln betrachtet werden, sondern nur die Entfernungen zwischen einem beliebigen und dem jeweils nächstgelegenen georäumlichen Artikel.

Die geringe Streuung der durchschnittlichen Pfadlängen impliziert, dass der Netzwerkansatz für die Bestimmung raumbezogener Informationen relativ unabhängig von der Größe und dem Entwicklungsstand einer spezifischen Wikipedia ist, unter der Annahme, dass der Anteil der georäumlichen Artikel konstant sei. Für Variationen dieses Anteils in Abhängigkeit verschiedener Sprachversionen und verschiedener Entwicklungsstadien von Wikipedia wird auf die Arbeiten von Dahinden (2011b) und Hecht & Moxley (2009) verwiesen. Eine durchgeführte Simulation hat gezeigt, dass der Einfluss eines niedrigeren Anteils an georäumlichen Artikeln auf die Verteilung der NGGR gering ist. Es kann also geschlussfolgert werden, dass die NGGR-Methode, auch für andere Sprachversionen als die verwendete deutsche, ähnliche Ergebnisse liefern müsste.

### 3 KOGNITIVER ANSATZ FÜR DIE BESTIMMUNG RAUMBEZOGENER INFORMATIONEN

Während der Netzwerkansatz für eine analytische Erfassung des Raumbezugs in Informationen geeignet scheint, ist er allein nicht dazu geeignet, eine Aussage darüber zu treffen, für welchen Anteil von Informationen der Raumbezug in der menschlichen Wahrnehmung relevant ist. Genauere Erkenntnisse hierüber sollte ein Experiment liefern, das darauf abzielte, die Bewertung des Raumbezugs durch Menschen zu untersuchen.

Da sich das menschliche Denken stark auf die Nutzung von Kategorien stützt, verwenden wir für den kognitiven Ansatz Kategorien des Georaumbezugs (KGR). In Anlehnung an Bill (2010, S. 12) teilen wir raumbezogene Informationen in die Kategorien „Direkter Georaumbezug“ (DirGR), „Indirekter Georaumbezug“ (IndGR) und „Kein Georaumbezug“ (KeinGR) ein. Diese Kategorien sind, im Gegensatz zum mathematisch begründeten Netzwerkgrad des Georaumbezugs, für Menschen fassbar. Da die ursprüngliche 80%-Zahl im Kontext geographischer Informationssysteme entstand, ist uns auch die Betonung von georäumlichen Informationen im Unterschied zu allgemein räumlichen Informati-

onen wichtig und wir verwenden deshalb bewusst den Begriff Georaumbezug. Die Abgrenzung georäumlicher Informationen ist eine Frage des Maßstabs, die z. B. in Haggett (2001, S. 37-39) dargestellt wird.

### Material

Für das Experiment wurden 1.100 Artikel aus der deutschen Wikipedia zufällig ausgewählt unter Beachtung zweier Randbedingungen:

1. Die Titel der Artikel sollten nicht mehrdeutig sein, z. B. Bank, um eine erhöhte Antwortstreuung durch Mehrdeutigkeiten zu vermeiden.
2. Die Artikel sollten möglichst bekannt sein, um einen hohen Anteil leicht kategorisierbarer Begriffe zu erzielen.

Die Bekanntheit von Artikeln wurde dabei mithilfe von Zugriffszahlen aus dem Zeitraum September 2010 bis August 2011 gemessen. Die Auswahl der Artikel wurde auf die 5 % meist zugriffenen Artikel in diesem Zeitraum beschränkt. Mithilfe der 1.100 ausgewählten Artikel wurden elf distinkte Sätze von je 100 Artikeln gebildet.

### Prozedur

Die Teilnahme am Experiment wurde über einen Onlinefragebogen realisiert. Der Fragebogen enthielt eine Einleitung, die die Motivation der Umfrage mit Bezug auf die 80 %-These darstellte. Die Teilnehmer bekamen keinen Hinweis darauf, dass in einer Parallelstudie (Netzwerkansatz) die expliziten Koordinaten von Wikipedia-Artikeln untersucht werden.

Zu Beginn wurden Angaben zur Person abgefragt: 1) Beruflicher Hintergrund oder Ausbildung in einer Geowissenschaft (ja/nein), 2) Kenntnis der 80 %-These (ja/nein), 3) Alter, 4) Geschlecht, 5) Student (ja/nein). Die Aufgabe der Teilnehmer im Hauptteil des Experimentes war es, Titel von Wikipedia-Artikeln einer der drei genannten Kategorien des Georaumbezugs zuzuordnen oder „Begriff unbekannt“ zu wählen. Die Definition des Begriffs Raumbezug nach Bollmann (2002) wurde hinsichtlich des Geoaspekts modifiziert (siehe: Hahmann 2011a) und den Probanden als Unterstützung vor Beginn der eigentlichen Aufgabe gezeigt. Nichtsdestotrotz wurden die Teilnehmer ermutigt, eigene Kriterien zur Kategorisierung zu nutzen. Die Artikel einer Gruppe wurden den Teilnehmern jeweils in zufälliger Reihenfol-

ge gezeigt, um systematische Fehler, z. B. durch Ermüdung der Teilnehmer, zu mindern. Abbildung 3 zeigt ein Beispiel der Zuordnungsaufgabe.

### Teilnehmer

Die Teilnehmerrekrutierung erfolgte über eine Pressemitteilung (Hahmann 2011b), E-Mails an Mitarbeiter und Studierende des Dresdner Instituts für Kartographie sowie an Bekannte der Autoren. Allen Teilnehmern wurde ein Startlink zur Verfügung gestellt, mit dem diese entsprechend eines Zufallsmechanismus in eine der Umfragegruppen weitergeleitet wurden.

## 4 ERGEBNISSE

### 4.1 ERGEBNISSE DES NETZWERKANSATZES

Basierend auf dem vollständigen Graphen der Wikipedia-Artikel konnte der Netzwerkgrad des Georaumbezugs für jeden Artikel berechnet werden. Die Ergebnisse sind in Tabelle 1 dargestellt. Der Anteil georäumlicher Artikel wird direkt durch NGGR 0 widerspiegelt. Es ist ersichtlich, dass nur NGGR 0 bis NGGR 3 relevant sind. Die 68 Artikel mit NGGR 4 können vernachlässigt werden. Artikel mit NGGR größer als 4 existieren nicht. Insge-

Abbildung 3: Beispiel der Zuordnungsaufgabe. Der Proband soll jeden Begriff in eine der drei Kategorien des Georaumbezugs einordnen oder „Begriff unbekannt“ wählen.

Das Ziel der Untersuchung war schließlich die Synthese von Netzwerkansatz und kognitivem Ansatz. Die Hypothesen waren dabei, dass 1. Artikel mit NGGR = 0 häufig als „direkter Georaumbezug“ kategorisiert werden, 2. Artikel mit NGGR > 0 selten als „direkter Georaumbezug“, dafür häufiger als „indirekter Georaumbezug“ kategorisiert werden, und dass 3. mit größerem NGGR der Anteil der Kategorisierung als „indirekter Georaumbezug“ zugunsten der Kategorisierung als „kein Georaumbezug“ sinkt.

samt sind 99,2 % aller Artikel mit mindestens einem georäumlichen Artikel verknüpft. Ein Rest von 0,8 % hat keine Verknüpfung zu einem georäumlichen Artikel. Es handelt sich dabei in der Regel um isolierte Artikel. Die kumulierten Anteile zeigen, dass Artikel mit NGGR 0 und NGGR 1 zusammen 78 % aller Wikipedia-Artikel ausmachen. NGGR-0-, NGGR-1- und NGGR-2-Artikel kommen zusammen sogar auf einen Anteil von 98,4 %.

Tabelle 2 zeigt einige prototypische Artikel für die drei relevantesten NGGR.

Bei NGGR 0 dominieren geographische Namen, wenngleich diese nicht exklusiv sind. Viele der Begriffe mit NGGR 1 sind Personen oder Begriffe, die eine starke räumliche Assoziation hervorrufen. Die meisten der NGGR-2-Begriffe sind abstrakte Konzepte. Beispielhaft sei ‚Carl XVI. Gustaf‘ als NGGR-1-Begriff genannt, der, wie die meisten anderen Personen in Wikipedia, durch direkte Links zu verschiedenen Orten zu einem NGGR-1-Begriff wird. Der NGGR-2-Artikel ‚Lineare Funktion‘ war zum Zeitpunkt der Untersuchung unter anderem mit dem Artikel ‚Äquator‘ verknüpft, der wiederum mit den georäumlichen Artikeln ‚Ecuador‘, ‚Kolumbien‘ und ‚Brasilien‘ verlinkt ist.

#### 4.2 DATEN DER EXPERIMENTEILNAHME

Die weite Verbreitung der 80%-These im Fachgebiet der Geowissenschaften wird dadurch deutlich, dass sie bei einem Anteil von 85 % der Probanden mit einem beruflichen Hintergrund oder einer Ausbildung in einer Geowissenschaft bekannt ist. Einschränkend sei hier gesagt, dass diese Zahl vermutlich durch eine höhere Motivation zur Experimenteildnahme durch Personen, die die These bereits kannten, beeinflusst sein mag. In der Gruppe der „Nicht-Geo-Teilnehmer“ ist die These jedoch nahezu unbekannt. Tabelle 3 zeigt die Daten zur Einordnung der Teilnehmer hinsichtlich ihrer Neutralität in Bezug auf die Forschungsfrage im Detail.

Insgesamt haben 380 Personen am Experiment teilgenommen. 33 % davon studierten zum Zeitpunkt der Teilnahme. Das Durchschnittsalter war 34,5 Jahre. Das Verhältnis Frauen zu Männern war 37 % zu 63 %.

#### 4.3 ERGEBNISSE DES KOGNITIVEN ANSATZES

Die Methode für die Analyse der Ergebnisse des kognitiven Ansatzes basiert auf

| NGGR | Anzahl der Artikel | Anteil  | Anteil, kumuliert |
|------|--------------------|---------|-------------------|
| 0    | 222.267            | 17,5 %  | 17,5 %            |
| 1    | 769.534            | 60,4 %  | 77,9 %            |
| 2    | 261.364            | 20,5 %  | 98,4 %            |
| 3    | 10.668             | 0,8 %   | 99,2 %            |
| 4    | 68                 | 0,0 %   | 99,2 %            |
| -1 * | 9.734              | 0,8 %   | 100,0 %           |
| Σ    | 1.273.635          | 100,0 % |                   |

\*Mit NGGR -1 werden isolierte Artikel bezeichnet.

**Tabelle 1:** Ergebnisse des Netzwerkansatzes. Der verwendete Wikipedia XML Dump ist vom 21.06.2011, die Koordinatendatei stammt vom 08.09.2011.

| NGGR 0         | NGGR 1                  | NGGR 2                |
|----------------|-------------------------|-----------------------|
| Straubing      | Carl XVI. Gustaf        | Lineare Funktion      |
| Zürich         | Friedrich Schiller      | Schießpulver          |
| Fränkische Alb | Französische Revolution | Selektive Wahrnehmung |
| Ulmer Münster  | Britische Teekultur     | Reportage             |
| Teldafax       | Aluminium               | Synthie Pop           |

**Tabelle 2:** Einige prototypische Beispiele für Artikel mit NGGR 0 bis NGGR 2

der Auswertung der Anteile der im Experiment zugewiesenen Kategorien pro Artikel. Abbildung 4 zeigt drei Beispielartikel mit unterschiedlichem NGGR. Der Begriff ‚Straubing‘ aus Gruppe 1 wurde beispielsweise durch 86 Teilnehmer bewertet. 73 davon wählten die Kategorie „Direkter Raumbezug“, weitere fünf wählten „Indirekter Raumbezug“ und acht wählten „Unbekannt“. Die Kategorie „Kein Raumbezug“ wurde von keinem Teilnehmer gewählt. Daraus ergeben sich die prozentualen Anteile von 84,9 % für DirGR, 5,8 % für IndGR, 0 % für KeinGR und 9,3 % für U. Im Weiteren wird angenommen, dass Teilnehmer, die U gewählt haben, sich mit der gleichen Wahrscheinlichkeit für eine Kategorie des Georaumbezugs entschieden hätten, wie alle anderen Teilnehmer, wenn sie den Begriff gekannt hätten. Demzufolge werden die Anteile der Kategorien relativ zur Men-

ge der nicht ‚U‘-Entscheidungen berechnet, woraus sich für das Beispiel Straubing die Anteile 93,6 % DirGR, 6,4 % IndGR und 0 % KeinGR ergeben. Die Ergebnisse aller 1.100 Artikel können online eingesehen werden (Hahmann 2012).

#### 4.4 SYNTHESE VON NETZWERKANSAZ UND KOGNITIVEM ANSAZ

Um den Zusammenhang zwischen Netzwerkansatz und kognitivem Ansatz untersuchen zu können, wird eine Synthese der Ergebnisse durchgeführt. Dafür werden für jede Kategorie des Georaumbezugs zweidimensionale Wertepaare aus dem NGGR eines Artikels und des jeweiligen relativen Anteils der entsprechenden Kategorie des Georaumbezugs gebildet. Für das Beispiel Straubing ergeben sich die drei Wertepaare 0/0,94 (NGGR und Anteil DirGR),

|               | Geo = ja<br>(abs.) | Geo = ja<br>(%) | Geo = nein<br>(abs.) | Geo = nein<br>(%) | Σ<br>(abs.) | Σ<br>(%) |
|---------------|--------------------|-----------------|----------------------|-------------------|-------------|----------|
| „80 %“ = ja   | 245                | 64,5            | 10                   | 2,6               | 255         | 67,1     |
| „80 %“ = nein | 47                 | 12,6            | 78                   | 20,5              | 125         | 32,9     |
| Σ             | 292                | 76,9            | 88                   | 23,1              | 380         | 100      |

**Tabelle 3:** Anteil der Teilnehmer mit einem beruflichen Hintergrund oder einer Ausbildung in einer Geowissenschaft (Geo = ja/nein) sowie Kenntnis der 80%-These („80 %“ = ja/nein)

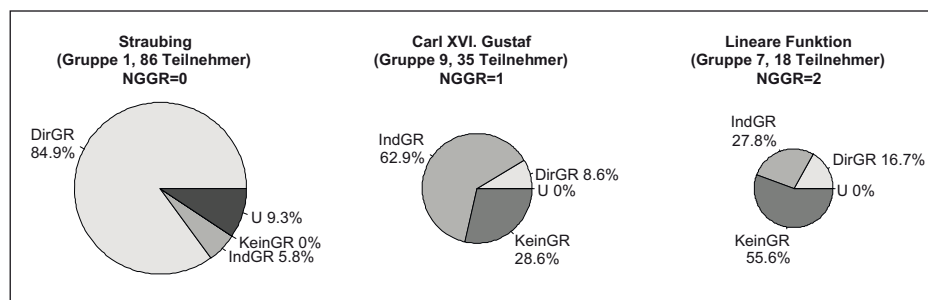


Abbildung 4: Ergebnisse der Kategorisierungsaufgabe. Die Größe der Kreise steht für die Anzahl der Teilnehmer.

1/0,06 (NGGR und Anteil IndGR) und 2/0 (NGGR und Anteil KeinGR). Für die drei Beispiele von Abbildung 4 ergeben sich somit insgesamt neun Wertepaare, die in Abbildung 5a dargestellt sind. Die Abbildungen 5b-5d zeigen den Zusammenhang zwischen Netzwerkansatz und kognitivem Ansatz für alle drei Kategorien des Georaumbezugs mithilfe der von Tukey (1977) eingeführten Box-Whisker-Plots.

Die Breite der Boxen ist in Relation zur Anzahl der jeweils umfassten Artikel skaliert. Die Höhe der Boxen stellt ein Maß für die Streuung der Ergebnisse dar. Um den Einfluss von vorhandenen Ausreißern zu minimieren, werden alle Mittelwerte als robuste Mittelwerte berechnet, mit einem Anteil von je 5 % gelöschten Werten ausgehend von beiden Enden der Verteilung (siehe Wilcox 2004, S. 56-59). Für die Berechnung der Standardabweichungen wird das ebenfalls robuste Schätzverfahren der winsorisierten Varianz verwendet (siehe Wilcox 2004, S. 62-63). Die Korrelationskoeffizienten werden nach Pearson berechnet.

Abbildung 5b zeigt den Zusammenhang zwischen NGGR und DirGR. Die Erwartung eines starken negativen Zusammenhangs bestätigt sich mit einem Korrelationskoeffizienten von  $-0,59$ . NGGR-0-Artikel wurden von den Teilnehmern zu 90 % als „Direkter Georaumbezug“ kategorisiert. Diese hohe Übereinstimmung bestätigt sowohl, dass die menschliche Wahrnehmung die Kategorie „Direkter Georaumbezug“ anwenden kann, als auch, dass die georäumlichen Artikel als solche erfasst werden. Die Voraussetzung für weitere Analysen mithilfe des Netzwerkansatzes ist somit erfüllt. Bei NGGR-2- und NGGR-3-Artikeln ist der Anteil an DirGR vernachlässigbar. Auch bei NGGR 1 ist er gering. Die oberen Ausreißer der NGGR-

1-Artikel sind mögliche Kandidaten, bei denen noch eine explizite Koordinate in den entsprechenden Wikipedia-Artikel eingefügt werden könnte.

Abbildung 5c zeigt den Zusammenhang von NGGR und indirektem Georaumbezug. Es ist ersichtlich, dass es sich um einen nicht linearen Zusammenhang mit einem Maximum bei NGGR 1 handelt. Der Anteil von IndGR für Artikel mit NGGR größer 1 nimmt der Anteil von Indirektem Georaumbezug stetig ab.

Der Zusammenhang von NGGR und KeinGR ähnelt dem inversen Zusammenhang von NGGR und DirGR. Hier gibt es

einen starken positiven Zusammenhang, bei dem gilt: Je höher der NGGR eines Artikels, desto höher die Wahrscheinlichkeit, dass dieser als KeinGR kategorisiert wird.

Aus der geringen Streuung der Antworten der NGGR-0-Artikel für alle drei KGR lässt sich ableiten, dass deren Einordnung relativ eindeutig war. Die relativ hohen Streuungen bei den Artikeln, die über einen (NGGR 1) oder zwei Links (NGGR 2) mit einem georäumlichen Artikel verknüpft sind, in den Kategorien IndGR und KeinGR zeigen, dass die Einordnung hier schwieriger war und dass die Abgrenzung zwischen IndGR und KeinGR unscharf ist.

#### 4.5 PRÄDIKTION DER KATEGORIEN DES GEORAUMBEZUGS FÜR DIE GESAMTE DEUTSCHE WIKIPEDIA

Der Anlass für die Forschungsarbeiten zu diesem Beitrag war die Suche nach einer Antwort auf die Frage, wie viel Prozent aller Informationen einen Georaumbezug haben. Hierfür haben wir mithilfe der Ergebnisse beider Ansätze eine Prädiktion für die gesamte deutsche Wikipedia berechnet. Tabelle 4 stellt die Ergebnisse dar.

Die Mittelwerte jeder NGGR-KGR-Kombination sind dabei die gleichen wie

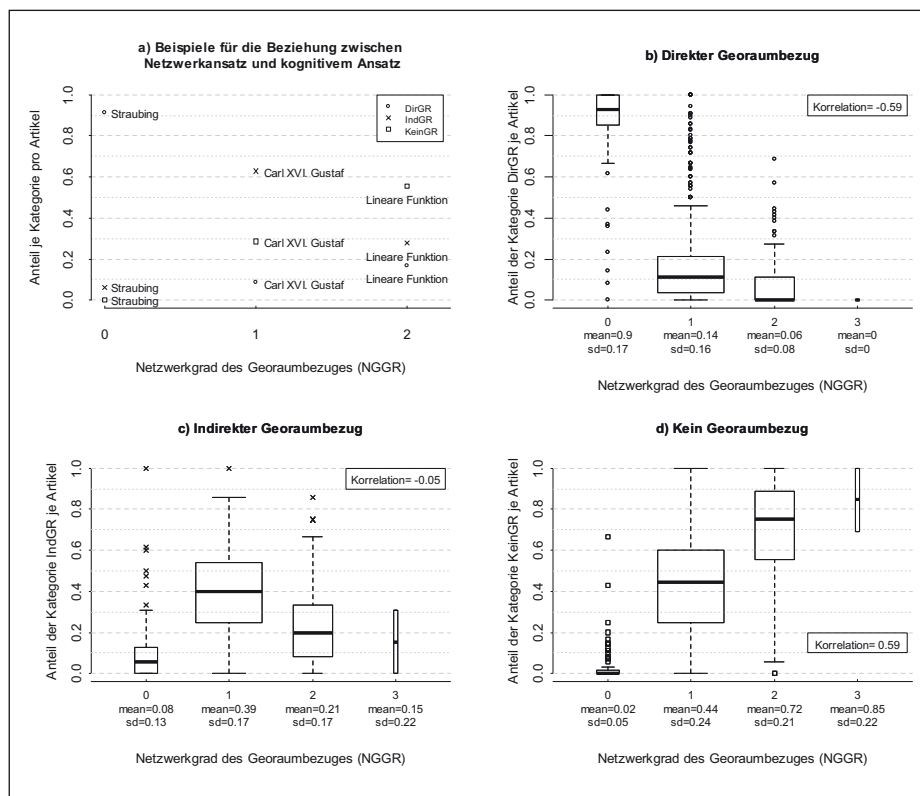


Abbildung 5: a) Die Anteile der Kategorien DirGR, IndGR und KeinGR und die NGGR der drei Beispielartikel (siehe Abbildung 4) auf ein X-Y Diagramm übertragen, b) – d) Box-Whisker-Diagramme für den Zusammenhang zwischen NGGR und verschiedenen KGR

| KGR    | NGGR | 0<br>17,6 %    | 1<br>60,9 %    | 2<br>20,7 %    | 3<br>0,8 %    | Σ      |
|--------|------|----------------|----------------|----------------|---------------|--------|
| DirGR  |      | 0,90 ± 0,04    | 0,14 ± 0,01    | 0,06 ± 0,01    | 0,00 ± 0      | 27.0 % |
|        |      | 15,2 % ± 0,7 % | 10,4 % ± 0,8 % | 1,4 % ± 0,2 %  | 0,0 % ± 0 %   | ±1.1 % |
| IndGR  |      | 0,08 ± 0,03    | 0,39 ± 0,01    | 0,21 ± 0,02    | 0,15 ± 0,34   | 30.3 % |
|        |      | 2,0 % ± 0,5 %  | 23,9 % ± 0,9 % | 4,4 % ± 0,5 %  | 0,1 % ± 0,3 % | ±1.2 % |
| KeinGR |      | 0,02 ± 0,01    | 0,44 ± 0,02    | 0,72 ± 0,03    | 0,85 ± 0,34   | 42.7 % |
|        |      | 0,7 % ± 0,2 %  | 26,9 % ± 1,2 % | 14,3 % ± 0,6 % | 0,8 % ± 0,3 % | ±1.4 % |

**Tabelle 4:** Hochrechnung der Anteile für die Kategorien des Georaumbezugs für die gesamte deutsche Wikipedia

in Abbildung 5b-5d. Die Standardfehler der Mittelwerte wurde mit dem robusten Verfahren nach Wilcoxon (2004, S. 63) mit einer Genauigkeitsforderung von  $2\sigma$  geschätzt. Die Anteile der verschiedenen NGGR sind angepasste Werte aus Tabelle 1, die relativ zu allen nicht isolierten Artikeln (NGGR = -1) berechnet wurden. NGGR 4 ist vernachlässigbar. Für jede NGGR-KGR-Kombination wird der Prozentwert des entsprechenden NGGR-Anteils mit dem Mittelwert der jeweiligen KGR multipliziert (Ergebnisse kursiv). Für den Standardfehler wird analog vorgegangen. Der Gesamtanteil jeder Kategorie berechnet sich aus der Summe der einzelnen Mittelwerte jeder Kategorie. Die Gesamtstandardfehler der Kategorien werden nach Gauß'schem Fehlerfortpflanzungsgesetz ermittelt.

Es ergibt sich also, dass für den direkten Georaumbezug ein Anteil von 27,0 % ( $\pm 1,1$  %), für den indirekten Georaumbezug ein Anteil von 30,3 % ( $\pm 1,2$  %) und für keinen Georaumbezug ein Anteil von 42,7 % ( $\pm 1,4$  %) hochgerechnet werden kann. Wird die Angabe der Ergebnisse in Form eines Intervalls bevorzugt, so können 26-28 % (DirGR), 29-32 % (IndGR) und 41-44 % (KeinGR) genannt werden.

## 5 DISKUSSION

Eine Einschränkung der vorgestellten Methode besteht darin, dass es nicht möglich ist, den Raumbezug auf der Ebene einzelner Instanzen zu bestimmen. Die im Experiment kategorisierten Titel von Wikipedia-Artikeln repräsentieren oft abstrakte Konzepte. Dabei ist es in den meisten Fällen weder mög-

lich von der Kategorisierung des abstrakten Konzepts auf die Kategorisierung konkreter Instanzen zu schließen, noch die Anzahl der Instanzen eines abstrakten Konzepts zu schätzen. So würde beispielsweise das abstrakte Konzept „Haus“ wohl eher nicht als „direkter Georaumbezug“ kategorisiert – im Gegensatz zu konkreten Instanzen, wie dem „Empire State Building“. Die Schätzung der Anzahl aller Häuser ist mithin nur schwer möglich. Darüber hinaus enthält Wikipedia abstrakte Konzepte wie beispielsweise „Philosophie“ oder „Synthie

Pop“, für die keine physischen Instanzen existieren. Für eine Untersuchung mit der gleichen Methode auf der Ebene von Instanzen könnten beispielsweise Datensätze des Semantischen Webs betrachtet werden, die es ermöglichen, einzelne Daten bzw. Instanzen mithilfe eines Graphen zu modellieren.

Ein weiteres Ziel zukünftiger Arbeiten könnte es sein, den Netzwerkansatz zur Quantifizierung des Georaumbezugs zu verfeinern. Abbildung 6 zeigt eine Wortwolke aus 200 zufällig ausgewählten



**Abbildung 6:** Wortwolken-Visualisierung der Georäumlichkeit von zufällig ausgewählten Begriffen aus der Untersuchung. Oben: Skalierung entsprechend des Netzwerkansatzes, unten: Visualisierung entsprechend der Ergebnisse des kognitiven Ansatzes.

Begriffen, die im Rahmen der Forschungsarbeiten mit beiden Ansätzen untersucht wurden. Die Größe der Wörter wurde in Abhängigkeit von NGGR und in Abhängigkeit der Anteile der Kategorien des Georaumbezugs skaliert. Größere dargestellte Begriffe stehen für räumlichere Begriffe. Man sieht, dass der Netzwerkansatz zwar sehr scharf klassifiziert, jedoch weniger feine Abstufungen hat, als die Ergebnisse der kognitiven Untersuchung. Eine Verbesserung des Netzwerkansatzes wäre möglich durch die Einführung von Kantengewichten auf der Basis der Anzahl oder auf der Basis der Semantik von Verlinkungen.

## 6 SCHLUSSFOLGERUNG

Die Forschungsarbeit, die in diesem Artikel dargestellt wurde, stellt einen Beitrag zur

Quantifizierung raumbezogener Informationen in nutzergenerierten Inhalten dar. Der vorgestellte Netzwerkansatz kann als Maß zur Abschätzung der „Georäumlichkeit“ von Informationen genutzt werden. Die geringen Standardfehler der Prädiktion (siehe Tabelle 4) stützen die Schlussfolgerung, dass die genutzte Kombination von Netzwerkansatz und kognitivem Ansatz dafür geeignet ist, den Anteil an Informationen mit Georaumbezug in Informationskorpora, die mithilfe der Graphentheorie modelliert werden können, zu bestimmen.

Die Ergebnisse der Studie zeigen, dass 57 % der Informationen in Wikipedia einen Georaumbezug haben. Die Zahl setzt sich aus 27 % Information mit direktem und 30 % Informationen mit indirektem Georaumbezug zusammen und steht einem Anteil von

43 % Informationen ohne Georaumbezug gegenüber. Diese Zahl ist zwar deutlich geringer als die ursprüngliche Annahme von 80 %, sie zeigt aber dennoch die Relevanz von georäumlichen Informationen. Aus Sicht der Autoren muss die Zahl deshalb nicht als Degradierung der Geowissenschaften gesehen werden, sondern als Chance, denn es ist sicher besser, eine niedrigere Zahl zu haben, die durch eine wissenschaftliche Untersuchung gestützt wird, als eine arbiträre höhere Zahl, die stets zweifelhaft erschien. Jedoch gilt auch für die Ergebnisse dieser Studie, dass sie nicht übergeneralisiert werden sollten. Eine Untersuchung mit kommunalen Datensätzen, auf die sich die ursprüngliche These bezog, hätte möglicherweise einen höheren Anteil an georäumlichen Informationen hervorgebracht.

## Literatur

Bill, R. (2010): Grundlagen der Geo-Informations-systeme. Heidelberg, Wichmann.

Bollmann, J. (2002): Raumbezug. In: Bollmann, J.; Koch, W. G. (Hrsg.): Lexikon der Kartographie und Geomatik. Spektrum, Heidelberg, S. 266.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; Wiener, J. (2000): Graph structure in the Web. In: Computer Networks, 33 (1-6), S. 309-320.

Capocci, A.; Servedio, V. D. P.; Colaiori, F.; Burrol, L. S.; Donato, D.; Leonardi, S.; Caldarelli, G. (2006): Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. In: Physical Review E, 74 (3).

Dahinden, T. (2011a): Wikipedia als Datengrundlage für Sachdaten von thematischen Karten. In: Kartographische Nachrichten, 61 (4), S. 177-183.

Dahinden, T. (2011b): Estimation of the Locations of the Language-Versions of Wikipedia – a Case Study on Geographic Data Mining. In: Ruas, A. (Ed.): Advances in Cartography and GIScience. Springer, S. 471-487.

Dispenser (2011): Geographic locations from Wikipedia articles. <http://toolserver.org/~dispenser/dumps/>, Zugriff 09/2011.

Fitzke, J.; Greve, K. (2010): Frei oder umsonst? – Nutzergenerierte Geoinformation zwischen Freiheit und Kostenlosigkeit. In: Strobl, J.; Blaschke, T.; Griesebner, G. (Hrsg.): Angewandte Geoinformatik 2010. Wichmann, Heidelberg, S. 732-741.

Haggett, P. (2004): Geographie. Eine globale Synthese. Ulmer, Stuttgart.

Hahmann, S. (2011a): Definition Begriff „Georaumbezug“. [http://kartographie.geo.tu-dresden.de/geospatial\\_experiment/definition.htm](http://kartographie.geo.tu-dresden.de/geospatial_experiment/definition.htm).

Hahmann, S. (2011b): Haben 80 % aller Daten einen Raumbezug??? – Dresdner Institut für Kartographie führt empirische Untersuchung durch. <http://www.geobranchen.de/index.php?option=content&task=view&id=5053>.

Hahmann, S. (2012): Experiment „Georaumbezug“. [http://kartographie.geo.tu-dresden.de/geospatial\\_experiment/results.htm](http://kartographie.geo.tu-dresden.de/geospatial_experiment/results.htm).

Hahmann, S.; Burghardt, D.; Weber, B. (2011): „80% of All Information is Geospatially Referenced“??? Towards a Research Framework:

Using the Semantic Web for (In)Validating this Famous Geo Assertion.

Hecht, B.; Moxley, E. (2009): Terabytes of Tobler: Evaluating the First Law in a Massive, Domain-Neutral Representation of World Knowledge. In: Hornsby, K. S.; Claramunt, C.; Denis, M.; Ligozat, G. (Eds.): Spatial Information Theory. LNCS, 5756. Springer, S. 88-105.

Tukey, J. W. (1977): Exploratory data analysis. Addison-Wesley.

Voss, J. (2005): Measuring Wikipedia. In: Ingwersen, P.; Larsen, B. (Eds.): Proceedings of the 10th International Conference on Scientometrics and Informetrics. International Society for Scientometrics and Informetrics.

Watts, D. J.; Strogatz, H. (1998): Collective dynamics of „small-world“ networks. In: Nature, 393, S. 440-442.

Wilcox, R. R. (2004): Introduction to robust estimation and hypothesis testing. Academic Press.

Zlatic, V.; Božičević, M.; Štefančić, H.; Domazet, M. (2006): Wikipedias: Collaborative web-based encyclopedias as complex networks. In: Physical Review E, 74 (1).