30 Jahre Agrarinformatik – Eine Textanalyse

Michael Clasen

Hochschule Hannover
Fakultät IV
Ricklinger Stadtweg 120
30459 Hannover
michael.clasen@hs-hannover.de

Abstract: In diesem Beitrag wird ein Überblick über die deutschsprachige Agrarinformatik der letzten 30 Jahre gegeben. Hierzu wurden nahezu alle Publikationen der Gesellschaft für Informatik in der Land-, Forst- und Ernährungswirtschaft (GIL) automatisiert nach auftretenden Wörtern und Autoren analysiert. Nicht digital vorliegende Texte wurden zunächst eingescannt und in Textform umgewandelt. Die Grundlage der Analyse bilden die GIL Tagungsbände sowie weitere Publikationen der GIL wie die Zeitschrift für Agrarinformatik (ZAI und eZAI) seit 1980. Insgesamt wurden 214333 unterschiedliche „Wörter“ aus 1998 Beiträgen in einer relationalen Datenbank erfasst und stehen somit für beliebige Auswertungen zur Verfügung. Dieser Beitrag schließt mit einigen Ergebnissen. Gerne steht der Autor für Forschungskooperationen bereit, in denen die Datenbasis bzgl. anderer Fragestellungen ausgewertet werden soll.

1 Einleitung

Bedrucktes Papier ist heutzutage kein knappes Gut mehr! Nicht die Menge an Informationen, sondern die Aufmerksamkeit des Lesers stellt den knappen Faktor dar. „So the complimentary scarce factor is the ability to understand that data and extract value from it” [Va09]. Der zielgerichteten Informationsverdichtung kommt also eine immer wichtigere Rolle zu. Beispielsweise fällt es nicht leicht zusammenzufassen, mit welchen Themen sich die deutschsprachige Agrarinformatik in den letzten 30 Jahren beschäftigt hat. Nicht weil die hierzu notwendigen Informationen nicht verfügbar wären; sie liegen jedoch häufig nicht in maschinenlesbarer Form vor und die schiere Menge an Informationen überfordert den menschlichen Geist, einen Überblick zu gewinnen.

Aus diesem Grunde wurde im WS2013/14 an der Hochschule Hannover ein Projekt zur Analyse sehr großer Textmengen gestartet und am Beispiel der Texte der deutschsprachigen Agrarinformatik durchgeführt [HH13]. Ziel des Projektes war es also 1) eine Methode und ein Werkzeug zur Analyse und zielgerichteten Aufbereitung sehr großer Textmengen zu entwickeln und 2) einen Überblick über die Themen- und Forschungsbereiche der deutschsprachigen Agrarinformatik zu geben. Im Folgenden wird zunächst die Analysemethode vorgestellt und anschließend werden erste Auswertungen zur Agrarin-

formatik vorgestellt. Da die Datenerhebung und -speicherung zunächst ohne Zielbezug erfolgt ist, kann die Datenbasis für diverse andere Forschungsfragen herangezogen werden, die sicherlich über die Agrarinformatik hinausgehen. Auch könnten die Methode und die entwickelten Werkzeuge dafür verwendet werden, weitere Textbestände zu speichern und zielgerichtet zu analysieren. Diesbezüglichen Forschungskooperationen steht der Autor offen gegenüber.

2 Analysemethode und –werkzeuge

Datenbasis

Die Datenbasis der Analyse bilden 23 Bände der Reihe „Informationsverarbeitung Agrarwissenschaft“ aus den Jahren 1980 (Band 1) bis 1993 (Band 25) [Re80]. Leider fehlen die Bände 2 und 4. Über Hinweise zu diesen Bänden wäre der Autor sehr dankbar. Des Weiteren wurden 15 Bände der „Berichte der Gesellschaft für Informatik in der Land-, Forst und Ernährungswirtschaft“ einbezogen, die überwiegend die Referate zu den GIL-Jahrestagungen 13- 24 aus den Jahren 1992 bis 2003 enthalten. In dieser Reihe fehlt leider der Band 2. Die Beiträge der Jahrestagungen 25 bis 33 sind in den Lecture Notes in Informatics der GI erschienen, die vollständig in die Analyse einbezogen worden sind. Schließlich wurden 109 Artikel aus 27 Ausgaben der Zeitschrift für Agrarinformatik (ZAI) der Jahrgänge 1999 bis 2005, sowie 32 Artikel aus den Jahrgängen 2006 bis 2009 der elektronischen Ausgabe der ZAI (eZAI) berücksichtigt. Die ZAI bzw. eZAI wurde also ab dem Jahrgang 1999 vollständig erfasst. Insgesamt wurden 1998 Beiträge aus 47 Bänden und 10 Jahrgängen der ZAI in die Analyse einbezogen. Die allermeisten Beiträge waren auf Deutsch, einige aber auch in englischer Sprache verfasst.

Vorgehensweise

Phase	Tätigkeit	Ergebnis
1	Einscannen nicht digital verfügbarer Tagungsbände	.pdf-Datei
2	OCR-Erkennung (Optical Character Recognition)	.txt-Datei
3	Erfassung der Text-Dateien in relationaler Datenbank mittels Java-Programm	Wörter in DB-Tabellen
4	Kategorisierung der einzelnen Wörter	Wörter in DB-Tabellen, kategorisiert
5	Fragespezifische Analysen per SQL	.xls-Datei
6	Fragespezifische Aufbereitung und Bereinigung der Daten	.xls-Datei
7	Fragespezifische Darstellung der Daten mittels Excel, Wordle, etc.	Tabellen oder Graphiken

Tabelle 1: Phasen der Textanalyse und Darstellung

Tabelle 1 gibt einen Überblick über die Vorgehensweise der Untersuchung sowie die jeweiligen Zwischenergebnisse. Da viele Beiträge nur in Buchform vorlagen, mussten zunächst ca. 30 Bände in ca. 180 Arbeitsstunden eingescannt werden. Hierzu wurde ein Epson Perfection V600 Scanner mit einer Auflösung von 600dpi verwendet. Ein ebenfalls verfügbarer Buchscanner mit einer Auflösung von lediglich 300dpi konnte nicht verwendet werden, da die Scann-Qualität für die OCR-Erkennung nicht ausreichend war.

Das Ergebnis des Scann-Prozesses war eine pdf-Datei, die mit der OCR-Erkennungssoftware „ABBYY Fine Reader Sprint Plus“ analysiert wurde und in eine Text-Datei umgewandelt werden konnte. Durch die geschickte Kombination des Fine Readers und des Adobe Acrobat Reader konnten anfängliche Probleme bei der Erkennung von Zeilenumbrüchen bei getrennten Wörter sowie durch unterschiedliche Rechtschreibungen weitgehend behoben werden. Die Textdateien wurden per cut&paste in ein eigens erstelltes Java-Programm übertragen, in dem die einzelnen Beiträge samt Informationen zu Buchreihe und Band erfasst und in eine relationale Datenbank geschrieben wurden. Das Ergebnis dieser ersten 3 Phasen ist eine relationale Datenbank mit 8 Tabellen, in denen gespeichert ist, welche Wörter in welchem Beitrag eines welchen Bandes auftreten. Zusammen mit Informationen zu den Bänden und Beiträgen wie z.B. Herausgeber und Autor, kann somit nahezu jede Analyse zum Auftreten von Wörtern und Autoren durchgeführt werden. Für die Agrarinformatik umfasst die Haupttabelle 214333 unterschiedliche Wörter bzw. sonstige Zeichenketten wie Abkürzungen oder Formeln. In Phase 4 wurden die 2500 häufigsten Wörter nach Wortarten oder agrarwissenschaftlichen Themenbereichen kategorisiert. Folgende Wortarten wurden bisher unterschieden: „Adjektiv-Adverb“, „Verb“, „Sonstiges“. Die Substantive wurden thematisch wie folgt kategorisiert: „Allgemeine IT-Begriffe“, „Agrarökonomie“, „Landtechnik“, „Bodenkunde“, „Pflanzenbau / Pflanzenzucht“, „Tierhaltung / Tierzucht“, „Wetter“, „Geoinformation“, „Agrar Allgemein“, „Orte“, „Personen“ und „sonstige Nomen“. Die Kategorisierung kann jederzeit mit dem Java-Tool erweitert oder verändert werden. Als problematisch hat sich jedoch erwiesen, dass eine Kategorisierung einzelner Wörter, also ohne Kontextbezug und Groß- und Kleinschreibung, nicht immer möglich ist. So könnte z.B. das Wort „werte“ ein „sonstiges Nomen“, ein „allgemeiner IT-Begriff“ oder auch ein Verb (ich werte etwas als…) sein.

Während die Phasen 1-4 nur einmal durchgeführt werden müssen, sind die Phasen 5-7 abfragespezifisch und müssen somit für jede Fragestellung erneut durchgeführt werden. Diese Vorgehensweise ermöglicht eine maximale Flexibilität und Kontextabhängigkeit bei Datenbereitstellung und -bereinigung sowie der Ergebnisdarstellung. So kann es beispielsweise für bestimmte Fragestellungen gleichgültig sein, ob die Abkürzung „PC“ oder das ausgeschriebene Wort „Personal Computer“ verwendet wurde. Hierzu kann einfach in Schritt 5 in der Excel-Tabelle eine Bereinigung der Daten durch Ersetzungsvorgänge in Excel durchgeführt werden. Will man aber vielleicht später analysieren, wann sich die Abkürzung „PC“ gegenüber der Langform „Personal Computer“ durchgesetzt hat, stehen diese Informationen weiterhin zur Verfügung. Voruntersuchungen haben darüber hinaus gezeigt, dass eine Datenbereinigung ohne Kenntnis des Analyseziels nicht möglich ist.

3 Analyseergebnisse der deutschsprachigen Agrarinformatik

Im Folgenden werden ausgewählte Ergebnisse der Untersuchung dargestellt. Abbildung 1 zeigt die Häufigkeit von Städte- und Ländernamen im Text der analysierten Beiträge. Demnach dominierten die Agrarhochburgen Weihenstephan, Hohenheim und Bonn die deutsche Agrarinformatikszene.

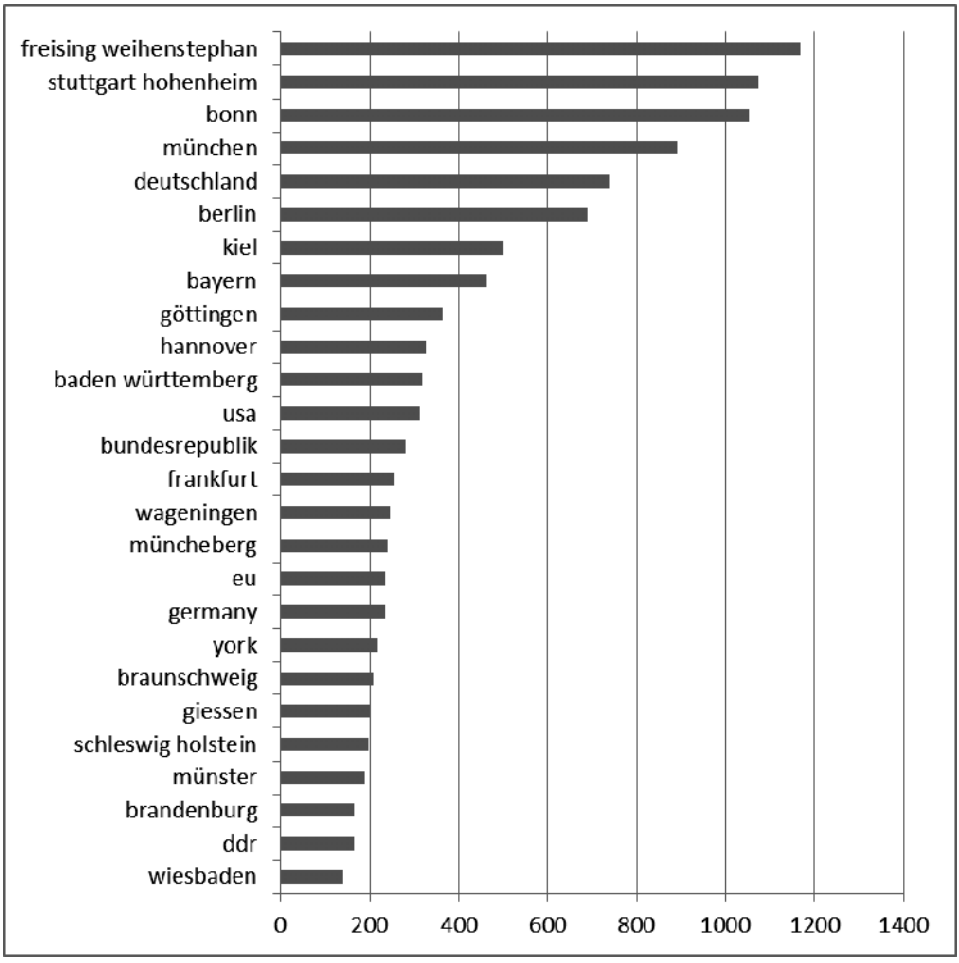

Abbildung 1: Häufigkeit von Städten und Ländern

Abbildung 2 ist ein Beispiel für eine Wortwolke, die mit dem Internetwerkzeug www.wordle.net erzeugt wurde. Sie zeigt die Häufigkeit von Begriffen der Kategorie „allgemeine IT“, wobei die Größe eines Wortes proportional zur Häufigkeit in der Datenbasis dargestellt wird. Dass die Begriffe „Daten“, „Agrarinformatik“ und „System“ dominieren, verwundert sicherlich nicht. Abbildung 3 stellt das Vorkommen der Begriffspaare „BTX“ und „Bildschirmtext“ sowie „Internet“ und „Web“ im Zeitverlauf dar. Es ist deutlich zusehen, dass sich die Agrarinformatik schon sehr früh mit der BTX-

Technologie beschäftigt hat, die erst 1983 in Deutschland flächendeckend eingeführt worden ist. Weiter ist zu sehen, dass das internetbasierte WorldWideWeb im Jahre 1994 den Bildschirmtext nahezu vollständig verdrängt hat.

Abbildung 2: Wortwolke zur Kategorie „allgemeine IT-Begriffe“ (Häufigkeit proportional zur Größe der Wörter)

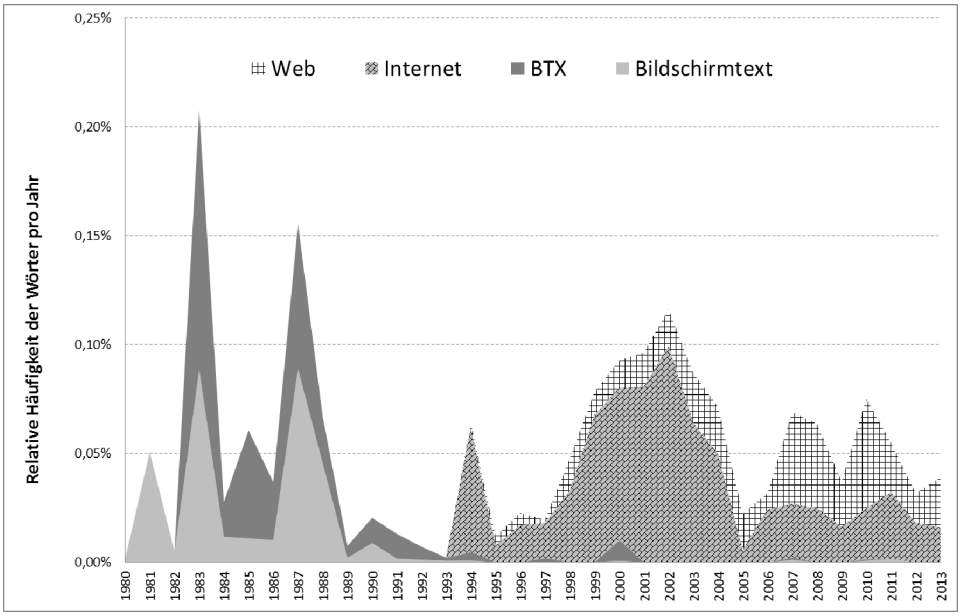

Abbildung 3: Relative Häufigkeit der Wörter Web, Internet, BTX und Bildschirmtext im Zeitverlauf; relativ zur Gesamtwortzahl des jeweiligen Jahres

Dies ist plausibel, da in diesem Jahre der kostenlose Webbrowser Netscape Navigator auf den Markt kam. Ein geringes Interesse an BTX erweckte vermutlich noch die Umbenennung auf Datex-J im Jahre 1997 und vergleichende Diskussionen im Rahmen des E-Business-Hypes um die Jahrtausendwende. Bzgl. der Webtechnologien kann man sehen, dass bis 2007 der Begriff „Internet“ das „Web“ dominierte; ab 2007 war es umgekehrt.

Agrar allgemein		Agrarökonomie		Tierhaltung / Tierzucht	
Wort	Häufigkeit	Wort	Häufigkeit	Wort	Häufigkeit
gil	1496	landwirtschaft	3191	tiere	545
landwirt	1441	unternehmen	2101	tier	351
landwirte	1056	dm	1932	futtermittel	266
düngung	945	kosten	1856	kuh	252
ertrag	761	management	1805	tierhaltung	229
crop	735	produktion	923	fütterung	228
umwelt	604	prozesse	818	ferkel	226
agrar	603	agriculture	767	sauen	219
standards	564	markt	758	animal	217
schlag	561	production	751	tierproduktion	216
nitrogen	544	ernährungswirtschaft	607	schweine	215
iso	536	schlagkartei	591	fleisch	212
standard	522	farming	591	ei	209
ernährung	496	preis	551	tieren	202

Landtechnik		Bodenkunde		Pflanzenbau / -zucht	
Wort	Häufigkeit	Wort	Häufigkeit	Wort	Häufigkeit
technik	695	boden	1313	pflanzen	735
maschinen	667	soil	1143	pflanzenschutz	680
geräte	418	böden	461	pflanzenbau	622
sensoren	362	bodens	382	winterweizen	619
sas	340	schicht	338	plant	616
sensor	272	bodenfeuchte	237	sorten	518
transponder	266	bodenart	214	stickstoff	495
ktbl	255	bodenschicht	209	pflanze	476
landtechnik	250	schichten	200	sorte	428
gerät	247	bodenbearbeitung	200	fruchtarten	423
maschine	234	acker	197	pflanzenproduktion	404
werkzeuge	212	grünland	196	getreide	365
anlage	211	soils	165	dünger	320
werkzeug	174	bodentemperatur	161	fruchtart	314

Tabelle 2: Häufigste Wörter pro Kategorie

Platz	Autor	Anzahl Beiträge	Platz	Autor	Anzahl Beiträge
1	Schiefer, G.	81	45	Precht, M.	9
2	Reiner, L.	52	46	Rößler, S.	9
3	Mangstl, A.	51	47	Spiller, A.	9
4	Spilke, J.	46	48	Übelhör, W.	9
5	Pohlmann, J.M.	44	49	Walther, P.	9
6	Doluschitz, R.	37	50	Bernhardt, H.	8
7	Theuvsen, L.	32	51	Krieter, J.	8
8	Petersen, B.	31	52	Lutze, G.	8
9	Engel, T.	28	53	Mückschel, C.	8
10	Wagner, P.	25	54	Odening, M.	8
11	Wenkel, K.-O.	25	55	Rosskopf, K.	8
12	Mirschel, W.	24	56	Schultz, A.	8
13	Bosch, J.	21	57	Selbeck, J.	8
14	Helbig, R.	20	58	Stricker, S.	8
15	Clasen, M.	19	59	Thiere, J.	8
16	Ohmayer, G.	19	60	Zickgraf, W.	8
17	Auernhammer, H.	18	61	Amon, H.	7
18	Penger, A.	17	62	Balmann, A.	7
19	Friedrich, H.	16	63	Bleiholder, H.	7
20	Fritz, M.	16	64	Dworak, V.	7
21	Müller, R.A.E.	16	65	Graeff, S.	7
22	Sundermeier, H.-H.	15	66	Haimerl, J.	7
23	Badewitz, S.	14	67	Hirschauer, N.	7
24	Franko, U.	14	68	Jungbluth, T.	7
25	Wendt, K.	14	69	Martini, D.	7
26	Wieland, R.	14	70	Nieschulze, J.	7
27	Giebler, P.	12	71	Poignee, O.	7
28	Kersebaum, K.C.	12	72	Priesack, E.	7
29	Hainzlmaier, J.	11	73	Distl, O.	6
30	Bergermeier, J.	10	74	Fröhlich, G.	6
31	König, E.	10	75	Gandorfer, M.	6
32	Mußhoff, O.	10	76	Geidel, H.	6
33	Noell, C.	10	77	Graf, R.	6
34	Schaaf, T.	10	78	Hannus, T.	6
35	Wendl, G.	10	79	Lex, J.	6
36	Claupein, W.	9	80	Piepho, H.-P.	6
37	Groeneveld, E.	9	81	Piotraschke, H.F.	6
38	Hahn, S.	9	82	Recke, G.	6
39	Hausen, T.	9	83	Rothmund, M.	6
40	Köhler, W.	9	84	Schmidt, F.	6
41	Kühbauch, W.	9	85	Schneider, M.	6
42	Kunisch, M.	9	86	Schreiner, H.	6
43	Mothes, V.	9	87	Streit, U.	6
44	Pitlik, L.	9	88	Wegehenkel, M.	6

Tabelle 3: Autoren mit mehr als 5 Beiträgen (inkl. Co-Autorenschaft)

Die jeweils häufigsten Wörter der Kategorien „Agrar allgemein“, „Agrarökonomie“, „Tierhaltung/Tierzucht“, „Landtechnik“, „Bodenkunde“ und „Pflanzenbau/-zucht“ sind in Tabelle 2 dargestellt. Es ist sicherlich plausibel, dass die zentralen Begriffe der jeweiligen Disziplinen am häufigsten auftreten; nämlich der Landwirt, das Unternehmen, die Tiere, Technik und Maschinen, Boden und Pflanzen. Neben diesen trivialen Erkenntnissen kann man aber auch erkennen, dass in der Agrarinformatik Kühe häufiger Forschungsgegenstand waren als Ferkel und Sauen. Im Pflanzenbau scheint der Pflanzenschutz für die Agrarinformatik eine herausragende Stellung einzunehmen.

Tabelle 3 zeigt schließlich alle Autoren, die in den untersuchten Texten mehr als 5 Beiträge platziert haben. Hierbei wurde nicht zwischen Einzel-Autorenschaft und Ko-Autorenschaft unterschieden.

Literaturverzeichnis und Danksagung

Alle analysierten Beiträge sind auf der Website der GIL www.gil.de unter Publikationen verfügbar.

[HH13] Mein Dank gilt Kevin Welzel, Thilo Pfalzgraf, Jan Siemer, Philipp Büntig, René Riedel, Barbara Halat, Muaz Malik und Oliver Cordes, die im WS 2013/14 im Rahmen eines Projektes der Hochschule Hannover die Tagungsbände eingescannt, die Werkzeuge entwickelt, die Datenbank aufgebaut und erste Analysen durchgeführt haben. Vielen Dank für die gute Arbeit.

[Re80] Für die Überlassung seiner 22 Exemplare der Reihe „Informationsverarbeitung Agrarwissenschaft“ aus den Jahren 1980 bis 1993 gilt mein besonderer Dank Herrn Kollegen Ludwig Reiner, Gründungsvorsitzender der GIL.

[Va09] Varian, H. (2009): Hal Varian on how the Web challenges managers, McKinsey & Company,
http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers