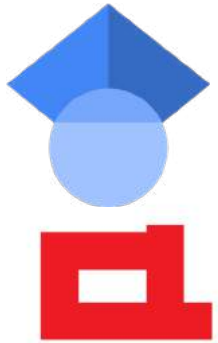
NLP Scholar: An Interactive Visual Explorer for Natural Language Processing Literature

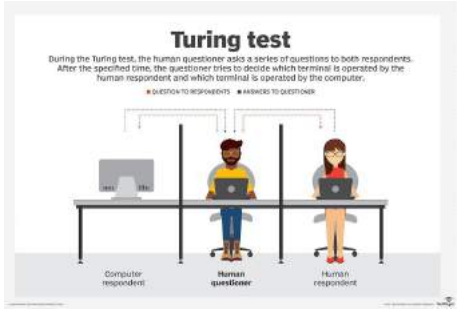


Saif M. Mohammad

Senior Research Scientist, National Research Council Canada

✉ Saif.Mohammad@nrc-cnrc.gc.ca @SaifMMohammad

Ten-Second History of NLP



- Turing Test proposed

1950

- Formation of the Association for Computational Linguistics (ACL)
  - originally named the Association for Machine Translation and Computational Linguistics (AMTCL)

1962

- First ACL conference

1965

- ACL, ELRA, and AFNLP publish peer-reviewed papers
  - journal articles and conference proceedings (long, short, demo, etc.)

- Broad interdisciplinary field
  - researchers from Computer Science, Linguistics, Information Science, Psychology, Social Sciences, Humanities, and more
  - thousands of papers published every year

Today

Need for a Search Engine Dedicated to NLP Literature

Desirable features:

- easy to use
- visual
- interactive
- captures impact of papers

Metrics of Research Impact (on subsequent scholarly work)

- Often derived from citations
  - number of citations, average citations, h-index, relative citation ratio, and impact factor (Bornmann and Daniel, 2009)
- However, citations do not always reflect quality or importance
  Impacted by:
  - systematic biases
  - atypical contributions
  - popularity of area
  - unethical practices (e.g. egregious self citations)

Nonetheless, given the lack of other easily applicable and effective metrics, citation metrics used as an imperfect but useful window into research impact

- often a factor in funding research and hiring scientists

The NLP Scholar Project

Data: extracted and aligned information from

- the ACL Anthology (AA)
- Google Scholar (GS)

to create a dataset of tens of thousands of NLP papers and their citations

NLP Scholar: A Dataset for Examining the State of NLP Research. LREC 2020.

Analyzing NLP Literature:

Examining Citations of Natural Language Processing Literature. ACL 2020.

Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations. ACL 2020.

Practical Tool:

NLP Scholar: An interactive visual explorer for this unified AA-GS dataset

- to find related work

Uses of NLP Scholar

- search for related work in various areas within NLP
- identify the highly cited articles on an interactive timeline
- identify past papers published in a venue of interest (such as ACL or COLING)
  - E.g., papers from ten years back published at ACL that have substantial citations
- examine changes in number of articles and number of citations in an area over time
- identify citation impact of different types of papers
  - e.g., short papers, shared task papers, demo papers, etc.

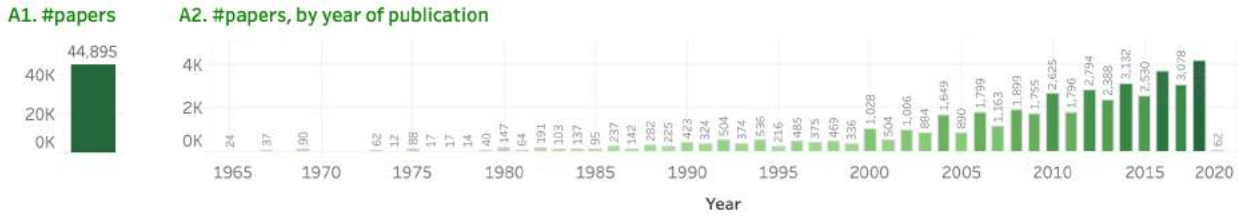
Other practical applications of the NLP Scholar dataset

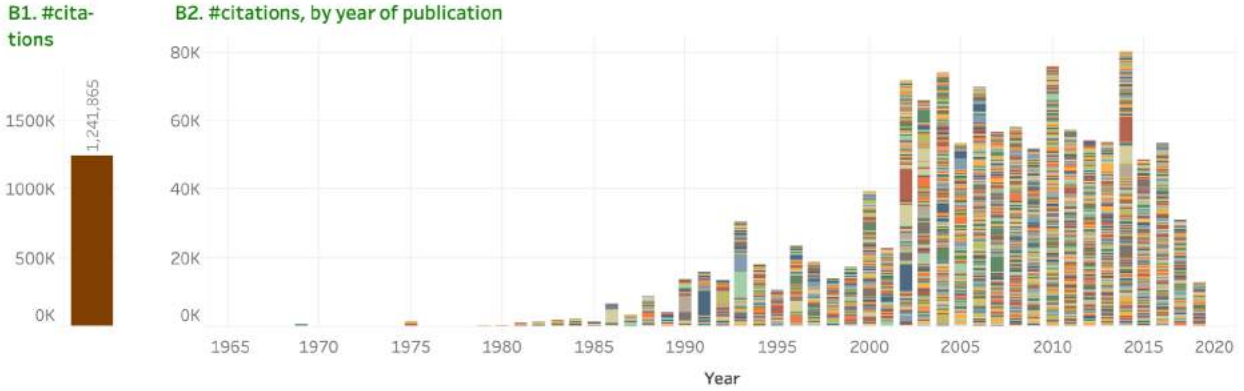
Web browser extensions that allow:

- users of GS to look up the aligned AA information
  - the full ACL BibTeX, poster, slides, access to proceedings from the same venue, etc.
- access from AA to the GS information of the aligned paper
  - number of citations, lists of papers citing the paper, etc.

NLP Scholar Dashboard

- A1: #papers
  - 44,895
- A2: #papers by year
  - substantial numbers since the 2000s
  - higher in alternate years (biennial conferences such as LREC, COLING
- B1: # citations
  - ~1.2 million citations (as of June 2019)
- B2: # citations by year
  - 2000s are the most notable





C. Papers

Row	Paper-Id	Paper-Title	Author(s)	Year	Url	#citations
1	P02-1040	Bleu: a Method for Automatic Evaluation of Machine Translation	Papineni, Kishore and Roukos, S..	2002	htt..	9,098
2	W02-1011	Thumbs up? Sentiment Classification using Machine Learning Techniques	Pang, Bo and Lee, Lillian and Vai..	2002	htt..	8,187
3	D14-1162	Glove: Global Vectors for Word Representation	Pennington, Jeffrey and Socher..	2014	htt..	7,965
4	J93-2004	Building a Large Annotated Corpus of English: The Penn Treebank	Marcus, Mitch and Santorini, Be..	1993	htt..	7,527
5	J91-4003	The Generative Lexicon	Pustejovsky, James	1991	htt..	6,593
6	P02-1053	Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsuper..	Turney, Peter	2002	htt..	5,642
7	D14-1179	Learning Phrase Representations using RNN Encoder-Decoder for Stati..	Cho, Kyunghyun and van Merrie..	2014	htt..	5,344
8	J93-2003	The Mathematics of Statistical Machine Translation: Parameter Estima..	Brown, Peter F. and Della Pietra..	1993	htt..	5,047
9	J90-1003	Word Association Norms, Mutual Information, and Lexicography	Church, Kenneth and Hanks, Pat..	1990	htt..	4,845
10	P07-2045	Moses: Open Source Toolkit for Statistical Machine Translation	Koehn, Philipp and Hoang, Hieu ..	2007	htt..	4,581

0K 5K 10K
#citations

D. Authors

Row	Author-name	#citations
1	Manning, Christoph..	54,587
2	Koehn, Philipp	19,412
3	Och, Franz Josef	18,620
4	Socher, Richard	17,506
5	Lee, Lillian	17,458
6	Jurafsky, Dan	16,405
7	Hovy, Eduard	16,292
8	Klein, Dan	15,881
9	Ney, Hermann	15,097
10	Dyer, Chris	14,745

0K 20K 40K 60K 80K
#citations

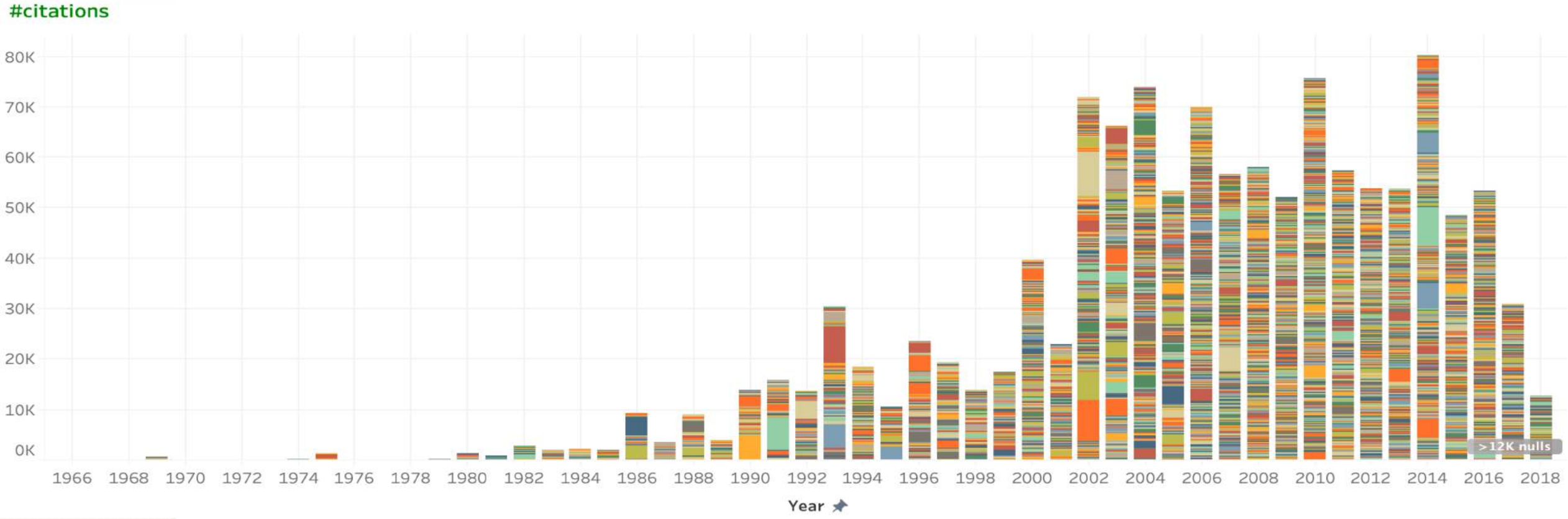
E. Search by year of publication, title term (unigram, bigram), or author name

Year of publication
1965 2019

Unigram | Bigram | Author Name

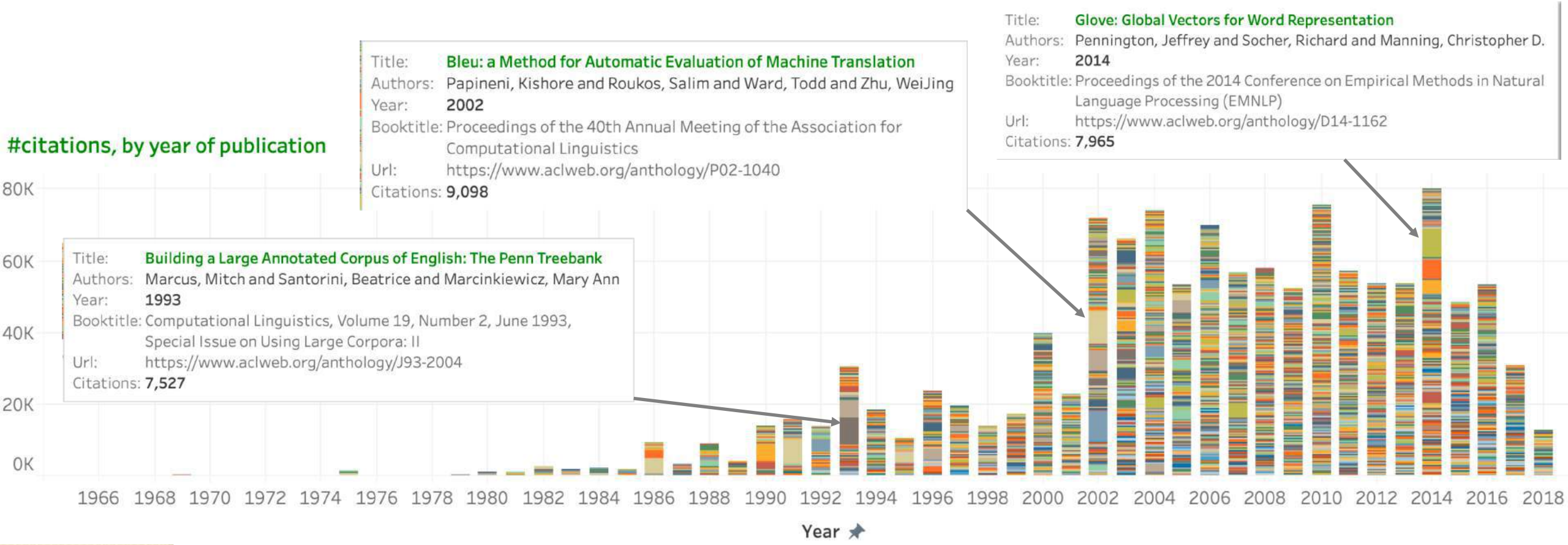
B2: # citations by year

- ~1.2 million citations (as of June 2019)
- colored segments correspond to each of the papers
- the height of a segment is proportional to the number of citations the paper has received

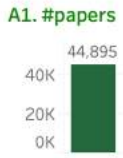


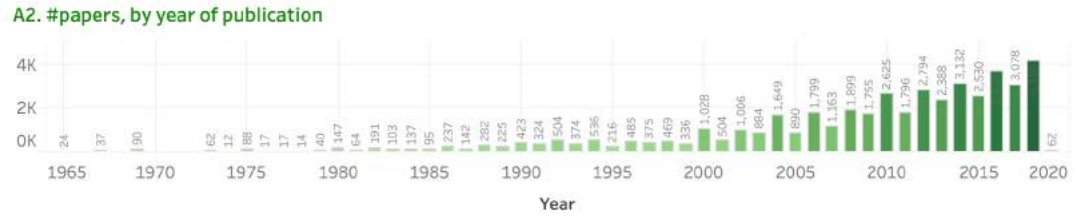
B2: # citations by year

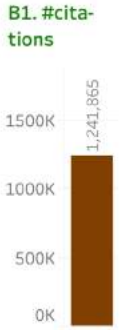
- ~1.2 million citations (as of June 2019)
- colored segments correspond to each of the papers
- the height of a segment is proportional to the number of citations the paper has received

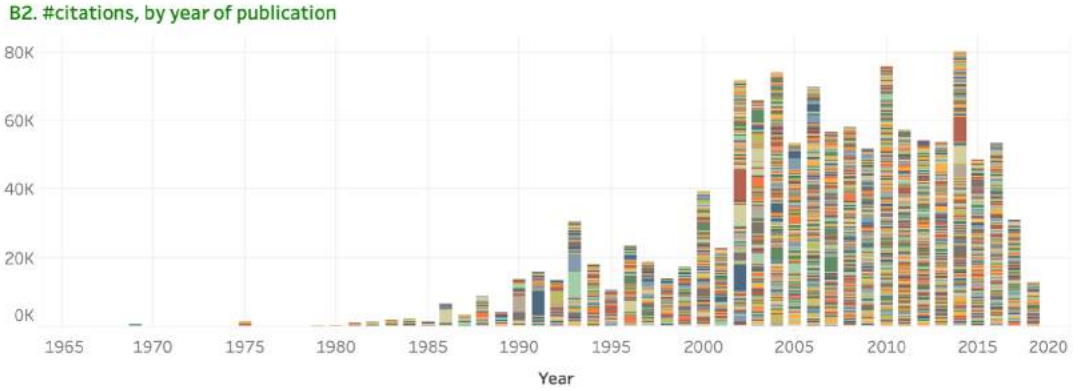


- A1: #papers
  - 44,895
- A2: #papers by year
  - substantial numbers since the 2000s
  - higher in alternate years (biennial conferences such as LREC, COLING
- B1: # citations
  - ~1.2 million citations (as of June 2019)
- B2: # citations by year
  - 2000s are the most notable
- C: most cited papers
  - *machine translation, sentiment analysis, word embeddings, syntax, semantics*
- D: most cited authors
- E: search options









C. Papers

Row	Paper-Id	Paper-Title	Author(s)	Year	Url	
1	P02-1040	Bleu: a Method for Automatic Evaluation of Machine Translation	Papineni, Kishore and Roukos, S..	2002	htt..	9,098
2	W02-1011	Thumbs up? Sentiment Classification using Machine Learning Techniques	Pang, Bo and Lee, Lillian and Vai..	2002	htt..	8,187
3	D14-1162	Glove: Global Vectors for Word Representation	Pennington, Jeffrey and Socher..	2014	htt..	7,965
4	J93-2004	Building a Large Annotated Corpus of English: The Penn Treebank	Marcus, Mitch and Santorini, Be..	1993	htt..	7,527
5	J91-4003	The Generative Lexicon	Pustejovsky, James	1991	htt..	6,593
6	P02-1053	Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsuper..	Turney, Peter	2002	htt..	5,642
7	D14-1179	Learning Phrase Representations using RNN Encoder-Decoder for Stati..	Cho, Kyunghyun and van Merrie..	2014	htt..	5,344
8	J93-2003	The Mathematics of Statistical Machine Translation: Parameter Estima..	Brown, Peter F. and Della Pietra..	1993	htt..	5,047
9	J90-1003	Word Association Norms, Mutual Information, and Lexicography	Church, Kenneth and Hanks, Pat..	1990	htt..	4,845
10	P07-2045	Moses: Open Source Toolkit for Statistical Machine Translation	Koehn, Philipp and Hoang, Hieu ..	2007	htt..	4,581

0K 5K 10K
#citations

D. Authors

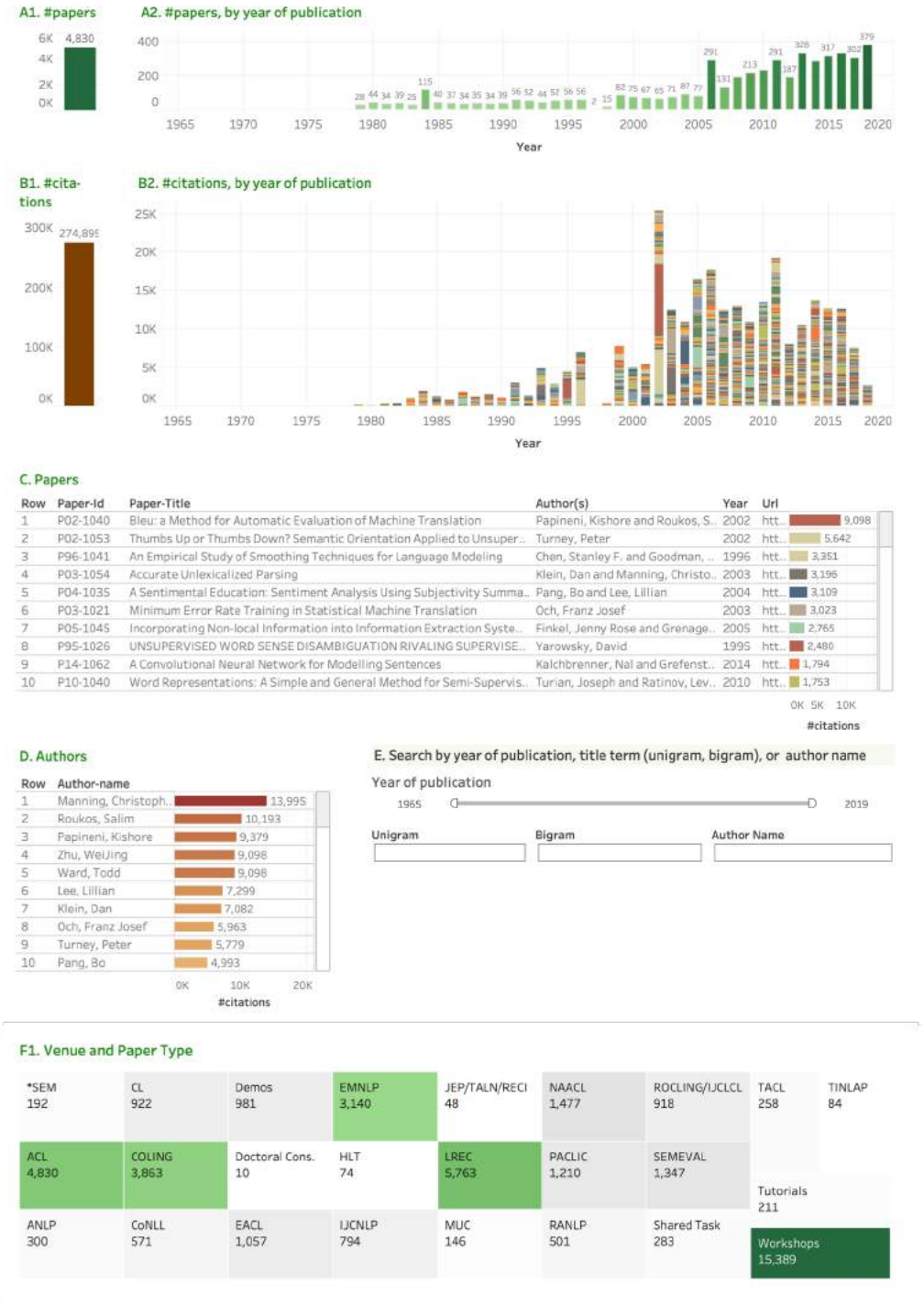
Row	Author-name	
1	Manning, Christoph..	54,587
2	Koehn, Philipp	19,412
3	Och, Franz Josef	18,620
4	Socher, Richard	17,506
5	Lee, Lillian	17,458
6	Jurafsky, Dan	16,405
7	Hovy, Eduard	16,292
8	Klein, Dan	15,881
9	Ney, Hermann	15,097
10	Dyer, Chris	14,745

0K 20K 40K 60K 80K
#citations

Four Other Dashboards

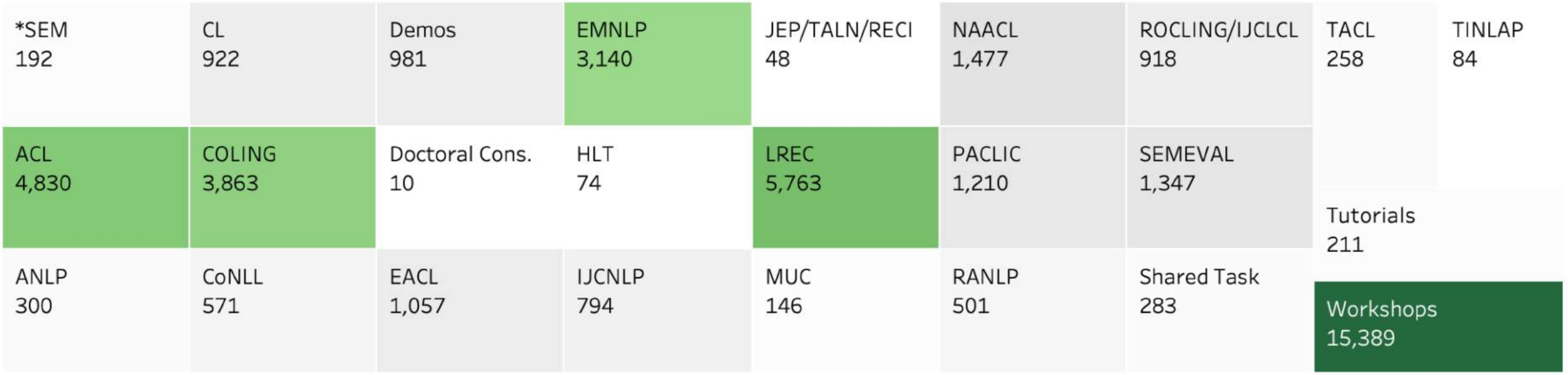
Main dashboard (A to E) + sixth element (F) for focused search

- F1: venues and paper types
- F2: title unigrams
- F3: title bigrams
- F4: language mentions in the title



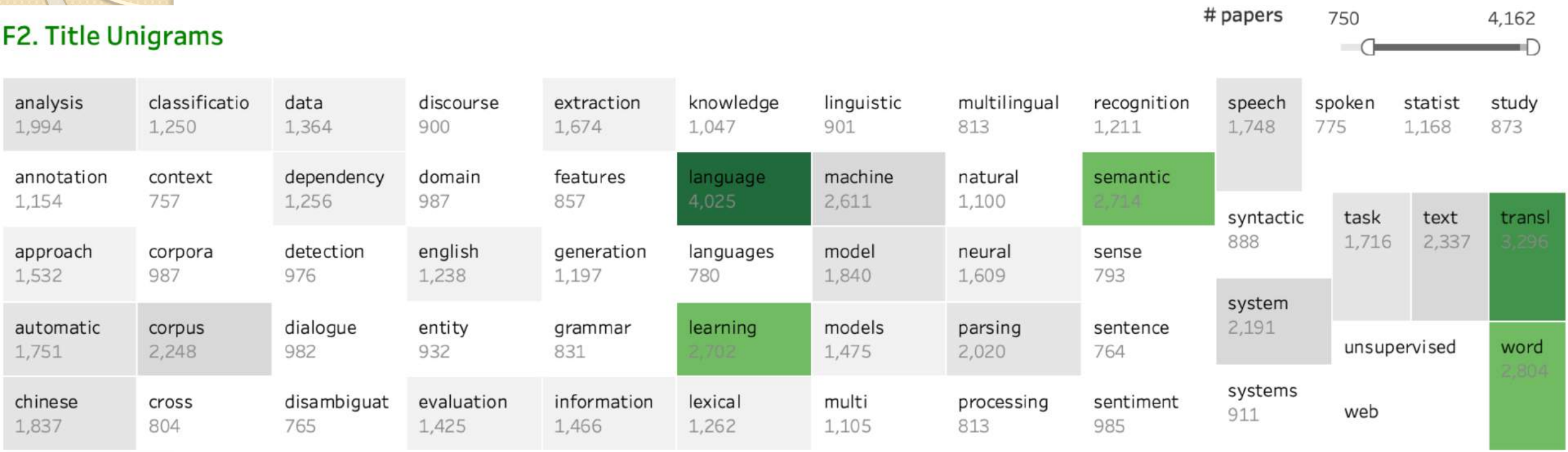
Treemap of Venues and Paper Types

F1. Venue and Paper Type



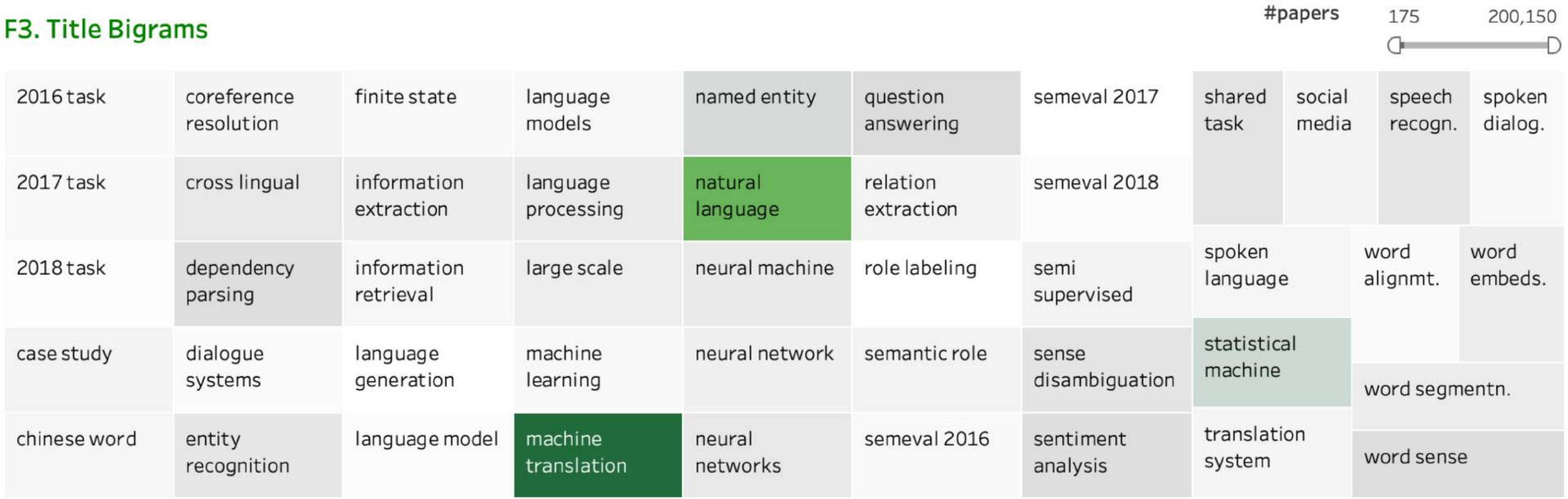
- Items are in alphabetic order
- Below each item is the number of papers
- Darker shades of green indicate higher numbers

Treemap of Common Title Unigrams



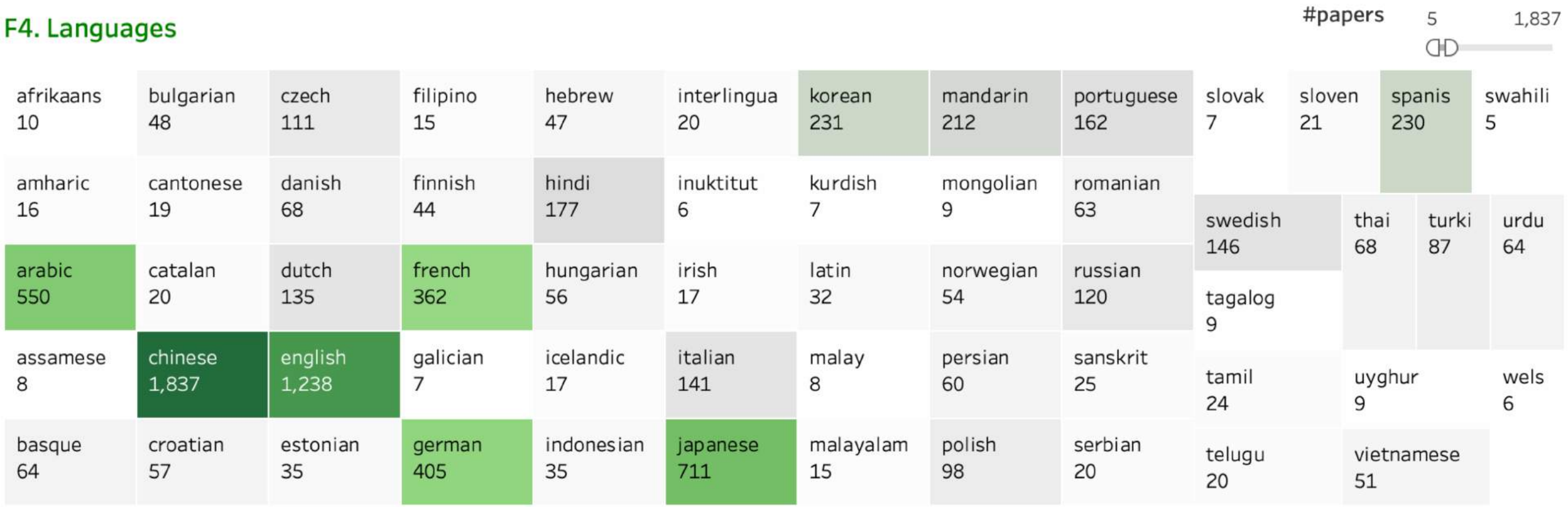
- Items are in alphabetic order
- Below each item is the number of papers
- Darker shades of green indicate higher numbers

Treemap of Common Title Bigrams



- Items are in alphabetic order
- Below each item is the number of papers
- Darker shades of green indicate higher numbers

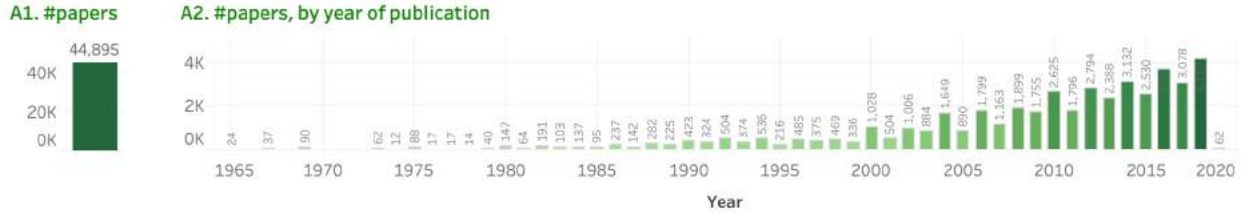
Treemap of Common Languages

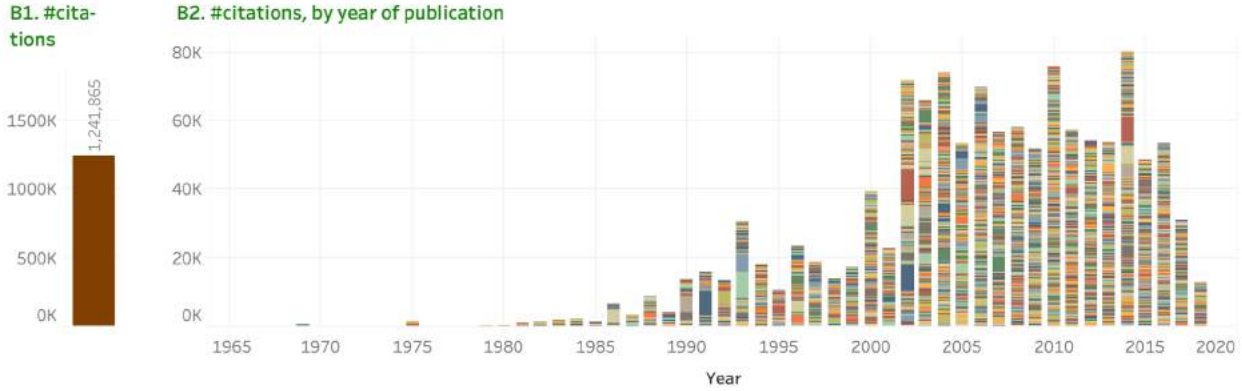


- Items are in alphabetic order
- Below each item is the number of papers
- Darker shades of green indicate higher numbers

NLP Scholar Dashboard

Main View





C. Papers

Row	Paper-Id	Paper-Title	Author(s)	Year	Url	#citations
1	P02-1040	Bleu: a Method for Automatic Evaluation of Machine Translation	Papineni, Kishore and Roukos, S..	2002	htt..	9,098
2	W02-1011	Thumbs up? Sentiment Classification using Machine Learning Techniques	Pang, Bo and Lee, Lillian and Vai..	2002	htt..	8,187
3	D14-1162	Glove: Global Vectors for Word Representation	Pennington, Jeffrey and Socher..	2014	htt..	7,965
4	J93-2004	Building a Large Annotated Corpus of English: The Penn Treebank	Marcus, Mitch and Santorini, Be..	1993	htt..	7,527
5	J91-4003	The Generative Lexicon	Pustejovsky, James	1991	htt..	6,593
6	P02-1053	Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsuper..	Turney, Peter	2002	htt..	5,642
7	D14-1179	Learning Phrase Representations using RNN Encoder–Decoder for Stati..	Cho, Kyunghyun and van Merrie..	2014	htt..	5,344
8	J93-2003	The Mathematics of Statistical Machine Translation: Parameter Estima..	Brown, Peter F. and Della Pietra..	1993	htt..	5,047
9	J90-1003	Word Association Norms, Mutual Information, and Lexicography	Church, Kenneth and Hanks, Pat..	1990	htt..	4,845
10	P07-2045	Moses: Open Source Toolkit for Statistical Machine Translation	Koehn, Philipp and Hoang, Hieu ..	2007	htt..	4,581

0K 5K 10K
#citations

D. Authors

Row	Author-name	#citations
1	Manning, Christoph..	54,587
2	Koehn, Philipp	19,412
3	Och, Franz Josef	18,620
4	Socher, Richard	17,506
5	Lee, Lillian	17,458
6	Jurafsky, Dan	16,405
7	Hovy, Eduard	16,292
8	Klein, Dan	15,881
9	Ney, Hermann	15,097
10	Dyer, Chris	14,745

0K 20K 40K 60K 80K
#citations

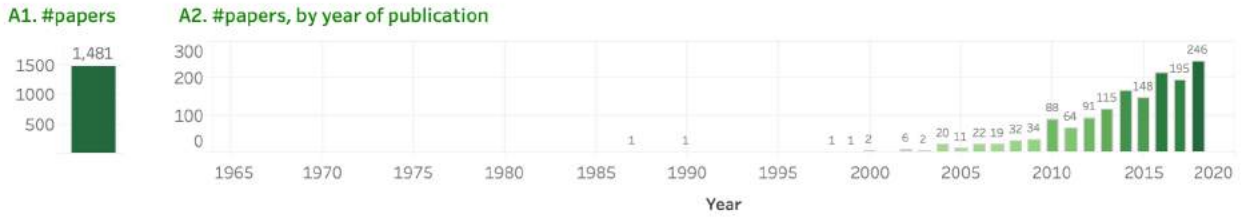
E. Search by year of publication, title term (unigram, bigram), or author name

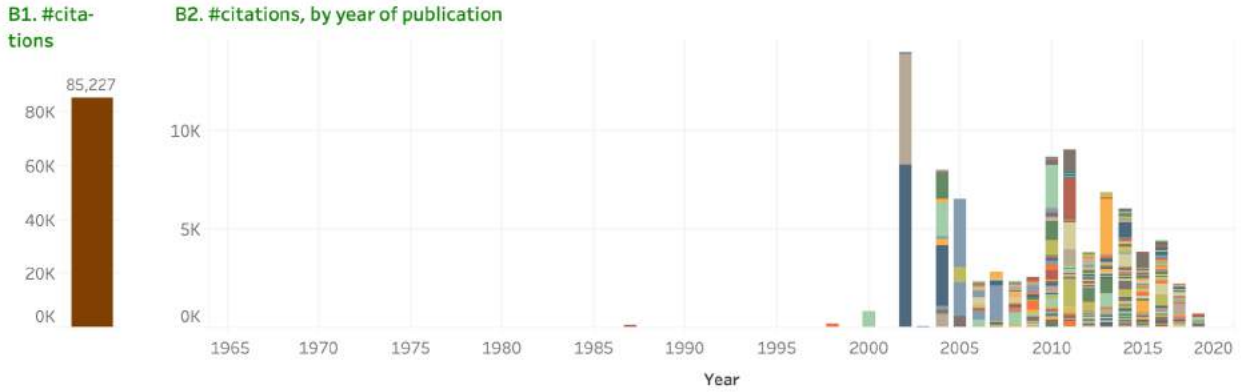
Year of publication
1965 — 2019

Unigram | Bigram | Author Name

NLP Scholar Dashboard

After entering terms associated with sentiment analysis in the Search Box.





C. Papers

Row	Paper-Id	Paper-Title	Author(s)	Year	Url	
1	W02-1011	Thumbs up? Sentiment Classification using Machine Learning Techniques	Pang, Bo and Lee, Lillian and Vai..	2002	htt..	8,187
2	P02-1053	Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsuper..	Turney, Peter	2002	htt..	5,642
3	H05-1044	Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis	Wilson, Theresa and Wiebe, Jan..	2005	htt..	3,487
4	P04-1035	A Sentimental Education: Sentiment Analysis Using Subjectivity Summa..	Pang, Bo and Lee, Lillian	2004	htt..	3,109
5	D13-1170	Recursive Deep Models for Semantic Compositionality Over a Sentimen..	Socher, Richard and Perelygin, ..	2013	htt..	2,798
6	L10-1-531	SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysi..	Baccianella, Stefano and Esuli, ..	2010	htt..	2,263
7	J11-2001	Lexicon-Based Methods for Sentiment Analysis	Taboada, Maite and Brooke, Jul..	2011	htt..	1,982
8	P05-1015	Seeing Stars: Exploiting Class Relationships for Sentiment Categorizati..	Pang, Bo and Lee, Lillian	2005	htt..	1,743
9	P07-1056	Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation ..	Blitzer, John and Dredze, Mark ..	2007	htt..	1,735
10	C04-1200	Determining the Sentiment of Opinions	Kim, SooMin and Hovy, Eduard	2004	htt..	1,723

0K 5K 10K
#citations

D. Authors

Row	Author-name	
1	Pang, Bo	13,039
2	Lee, Lillian	13,039
3	Vaithyanathan, Shi..	8,187
4	Turney, Peter	6,148
5	Ng, Andrew Y.	5,158
6	Manning, Christoph..	4,518
7	Wilson, Theresa	4,502
8	Wiebe, Janyce	4,409
9	Potts, Christopher	4,257
10	Socher, Richard	3,823

0K 5K 10K 15K
#citations

E. Search by year of publication, title term (unigram, bigram), or author name

Year of publication
1965 to 2019

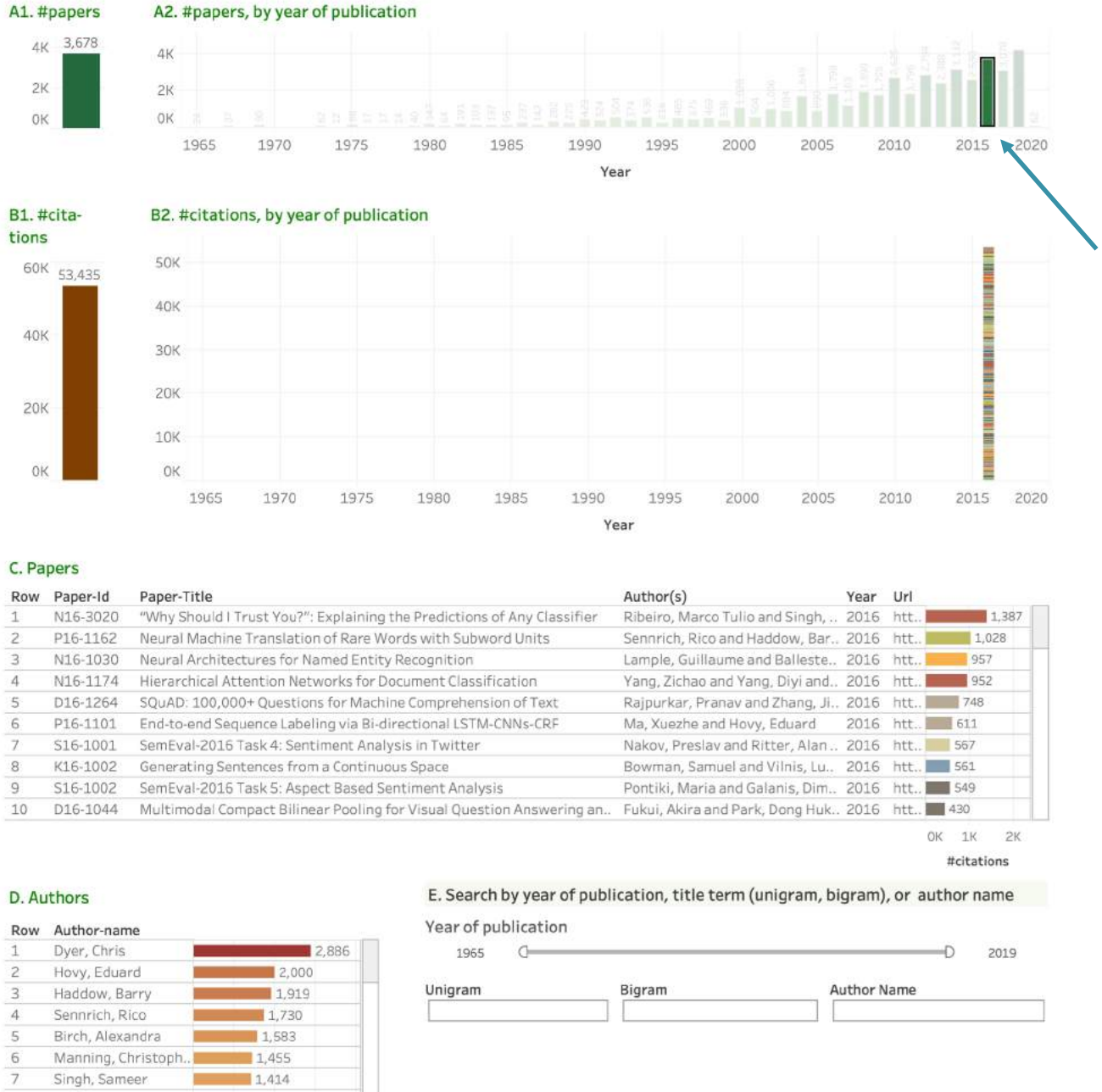
Unigram
#emotional
affect
emotion
emotional
emotions
orientation
sentiment
stance
valence

Bigram

Author Name

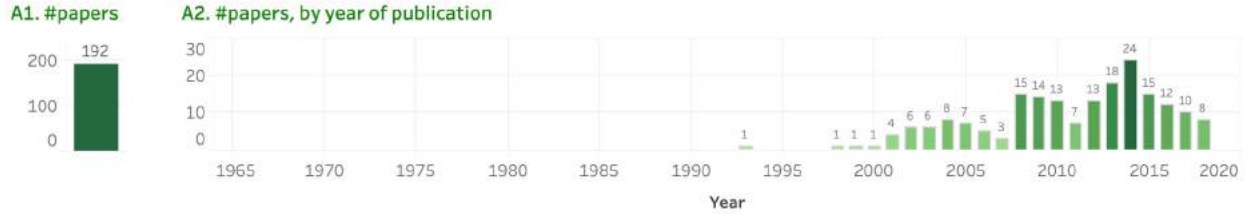
NLP Scholar Dashboard

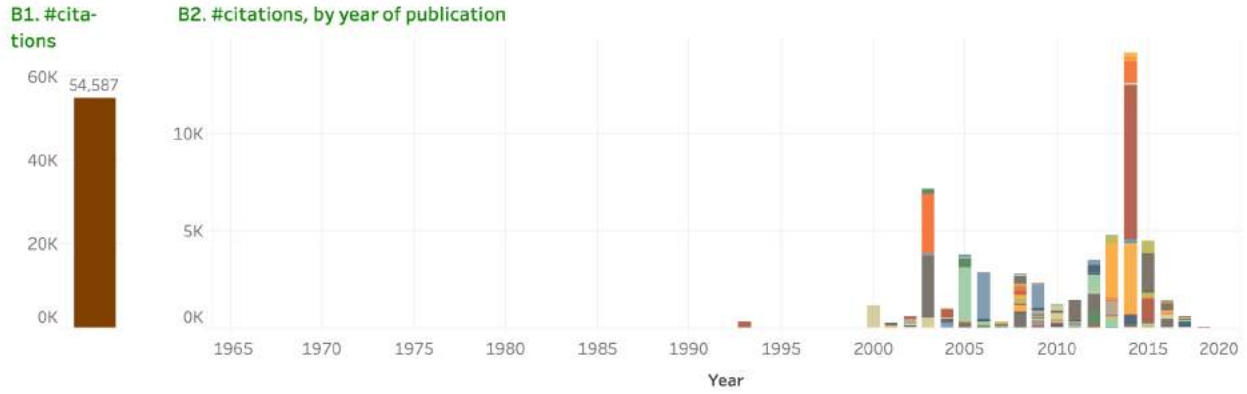
Clicking on the 2016 bar
in the #papers by year viz (A2)



NLP Scholar Dashboard

Clicking on ‘Manning, Christopher’ in the Authors List (D)





C. Papers

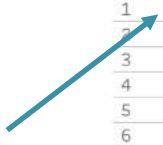
Row	Paper-Id	Paper-Title	Author(s)	Year	Url	
1	D14-1162	Glove: Global Vectors for Word Representation	Pennington, Jeffrey and Socher..	2014	htt..	7,965
2	P14-5010	The Stanford CoreNLP Natural Language Processing Toolkit	Manning, Christopher D. and Su..	2014	htt..	3,543
3	P03-1054	Accurate Unlexicalized Parsing	Klein, Dan and Manning, Christo..	2003	htt..	3,196
4	N03-1033	Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network	Toutanova, Kristina and Klein, D..	2003	htt..	3,083
5	D13-1170	Recursive Deep Models for Semantic Compositionality Over a Sentimen..	Socher, Richard and Perelygin, ..	2013	htt..	2,798
6	P05-1045	Incorporating Non-local Information into Information Extraction Syste..	Finkel, Jenny Rose and Grenage..	2005	htt..	2,765
7	L06-1-260	Generating Typed Dependency Parses from Phrase Structure Parses	de Marneffe, MarieCatherine a..	2006	htt..	2,414
8	D15-1166	Effective Approaches to Attention-based Neural Machine Translation	Luong, Minh-Thang and Pham, ..	2015	htt..	1,961
9	D09-1026	Labeled LDA: A supervised topic model for credit attribution in multi-lab..	Ramage, Daniel and Hall, David ..	2009	htt..	1,168
10	W00-1308	Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-S..	Toutanvoa, Kristina and Mannin..	2000	htt..	1,164

0K 5K 10K
#citations

D. Authors

Row	Author-name	
1	Manning, Christoph..	54,587
2	Koehn, Philipp	19,412
3	Och, Franz Josef	18,520
4	Socher, Richard	17,506
5	Lee, Lillian	17,458
6	Jurafsky, Dan	16,405
7	Hovy, Eduard	16,292
8	Klein, Dan	15,881
9	Ney, Hermann	15,097
10	Dyer, Chris	14,745

0K 20K 40K 60K 80K
#citations



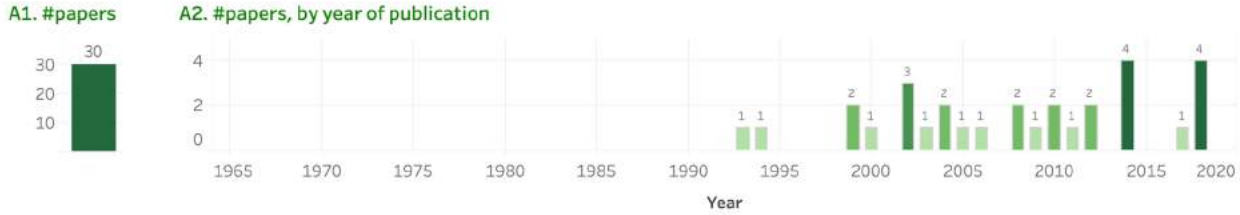
E. Search by year of publication, title term (unigram, bigram), or author name

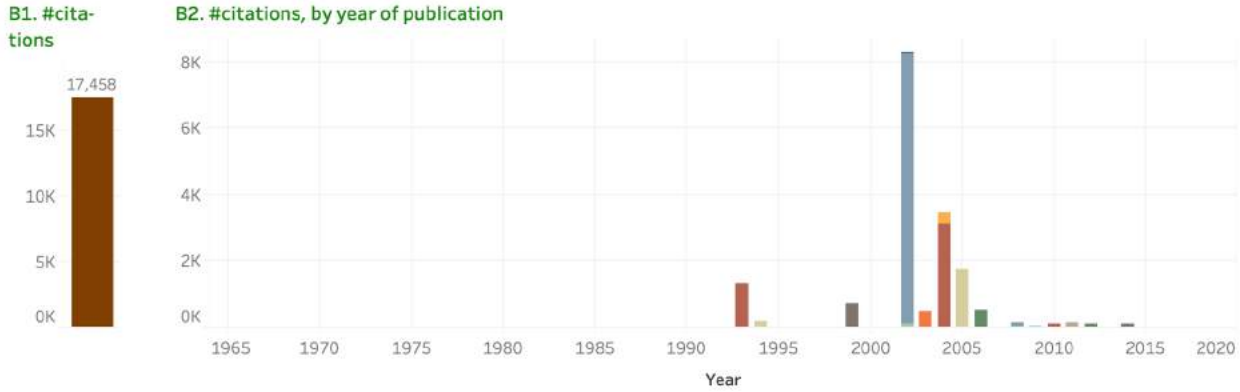
Year of publication
1965 2019

Unigram | Bigram | Author Name

NLP Scholar Dashboard

Clicking on 'Lee, Lillian' in the Authors List (D)





C. Papers

Row	Paper-Id	Paper-Title	Author(s)	Year	Url	
1	W02-1011	Thumbs up? Sentiment Classification using Machine Learning Techniques	Pang, Bo and Lee, Lillian and Vai..	2002	htt..	8,187
2	P04-1035	A Sentimental Education: Sentiment Analysis Using Subjectivity Summa..	Pang, Bo and Lee, Lillian	2004	htt..	3,109
3	P05-1015	Seeing Stars: Exploiting Class Relationships for Sentiment Categorizati..	Pang, Bo and Lee, Lillian	2005	htt..	1,743
4	P93-1024	DISTRIBUTIONAL CLUSTERING OF ENGLISH WORDS	Pereira, Fernando and Tishby, N..	1993	htt..	1,322
5	P99-1004	Measures of Distributional Similarity	Lee, Lillian	1999	htt..	697
6	W06-1639	Get out the vote: Determining support or opposition from Congressiona..	Thomas, Matt and Pang, Bo and..	2006	htt..	536
7	N03-1003	Learning to Paraphrase: An Unsupervised Approach Using Multiple-Seq..	Barzilay, Regina and Lee, Lillian	2003	htt..	504
8	N04-1015	Catching the Drift: Probabilistic Content Models, with Applications to G..	Barzilay, Regina and Lee, Lillian	2004	htt..	341
9	P94-1038	Similarity-Based Estimation of Word Cooccurrence Probabilities	Dagan, Ido and Pereira, Fernan..	1994	htt..	183
10	W11-0609	Chameleons in Imagined Conversations: A New Approach to Understand..	DanescuNiculescuMizil, Cristia..	2011	htt..	147

0K 5K 10K
#citations

D. Authors

Row	Author-name	
1	Manning, Christoph..	54,587
2	Koehn, Philipp	19,412
3	Och, Franz Josef	18,520
4	Socher, Richard	17,506
5	Lee, Lillian	17,458
6	Jurafsky, Dan	16,405
7	Hovy, Eduard	16,292
8	Klein, Dan	15,881
9	Ney, Hermann	15,097
10	Dyer, Chris	14,745

0K 20K 40K 60K 80K
#citations

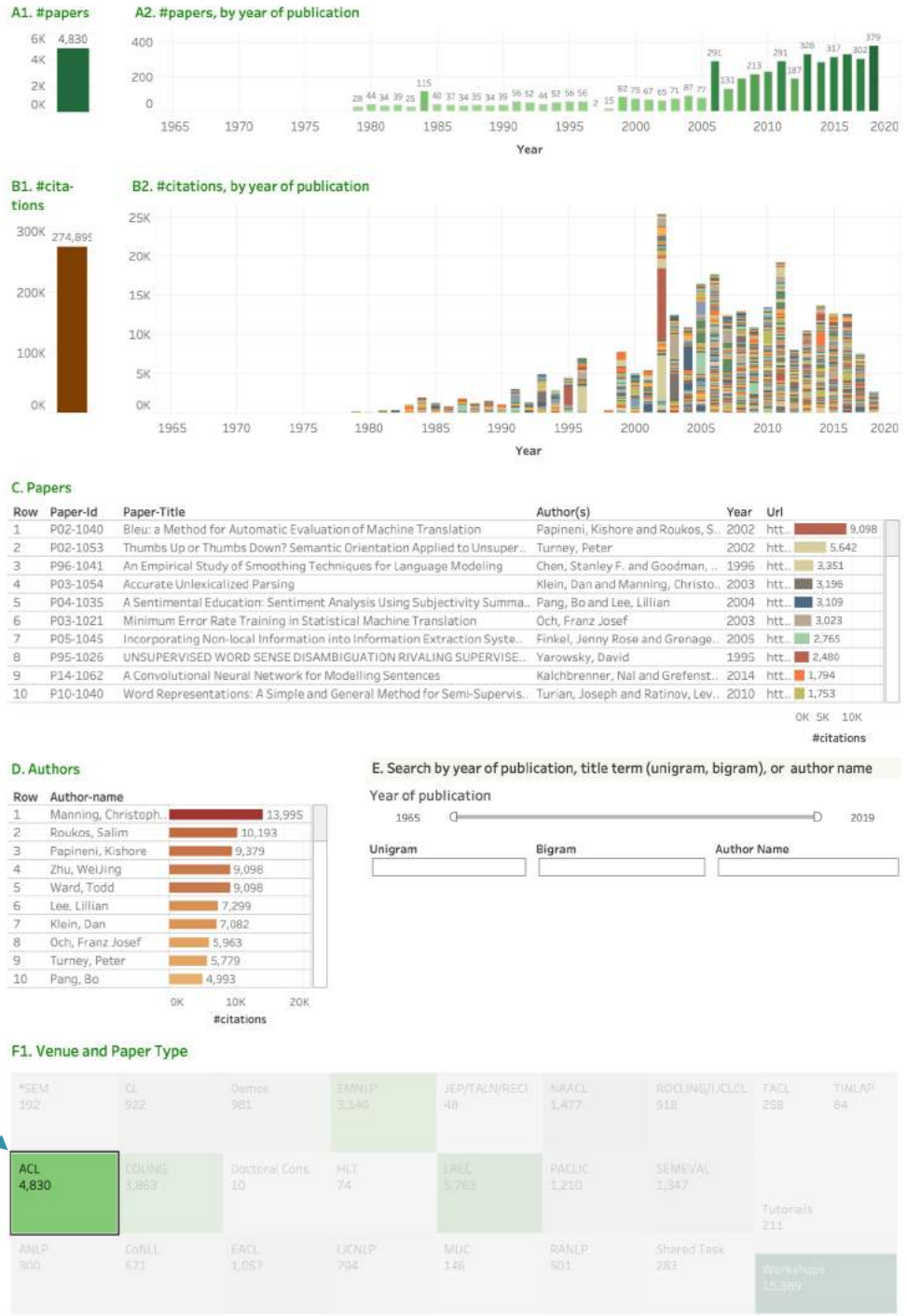
E. Search by year of publication, title term (unigram, bigram), or author name

Year of publication
1965 2019

Unigram | Bigram | Author Name

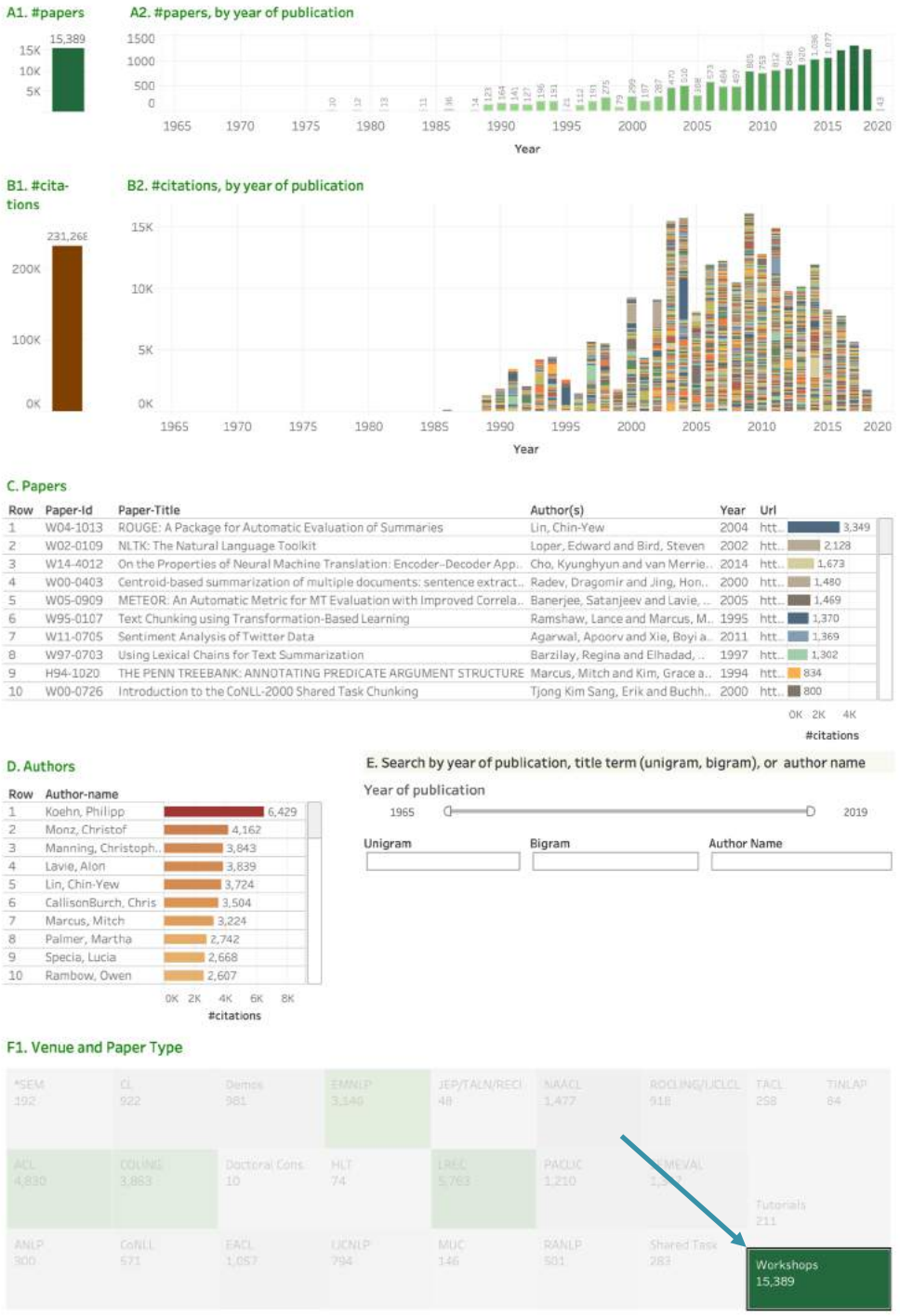
NLP Scholar Dashboard

Clicking on 'ACL' in the
Venue and Paper Type treemap (F1)



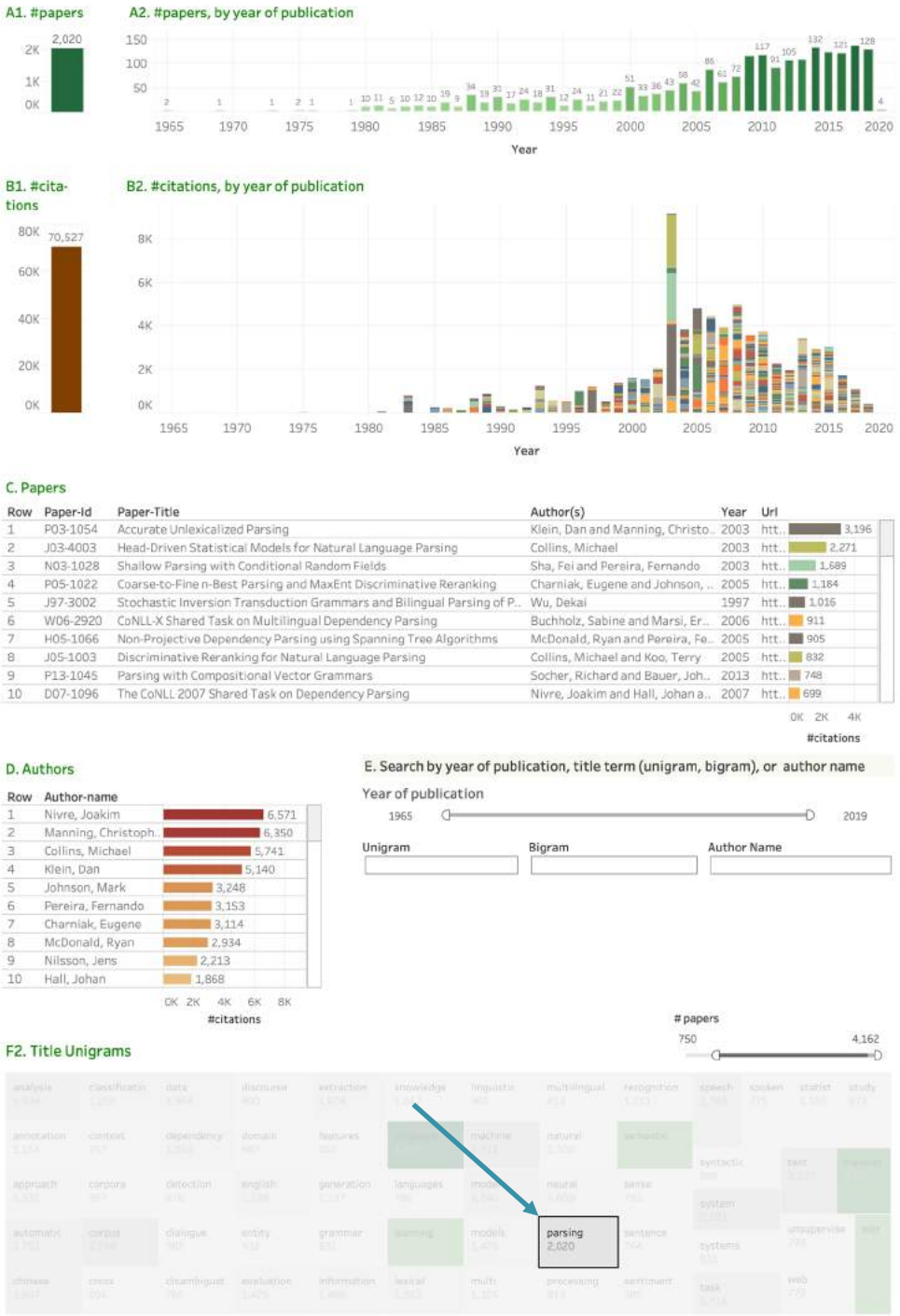
NLP Scholar Dashboard

Clicking on 'Workshops' in the
Venue and Paper Type treemap (F1)



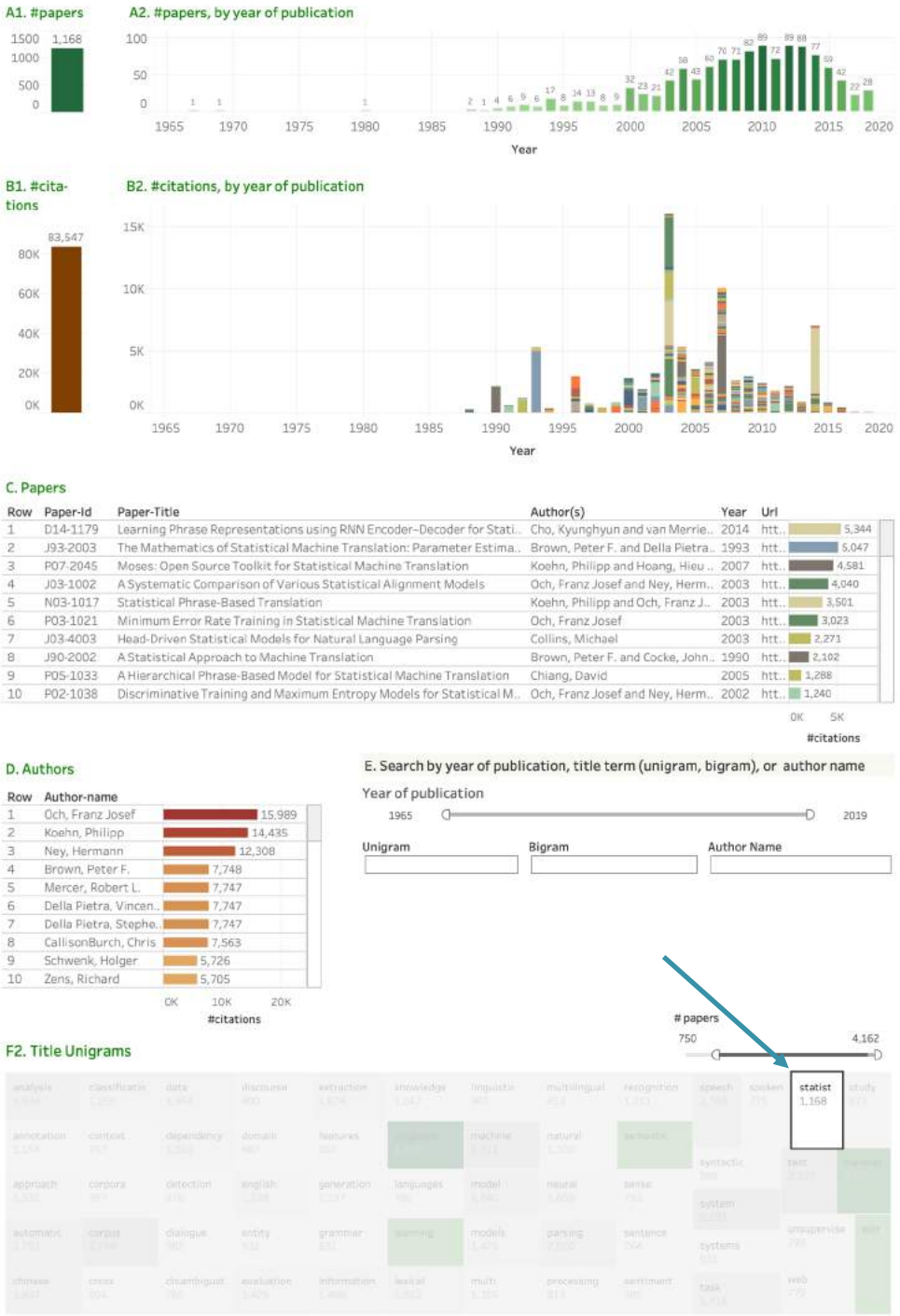
NLP Scholar Dashboard

Clicking on ‘parsing’ in the
Title Unigrams treemap (F2)



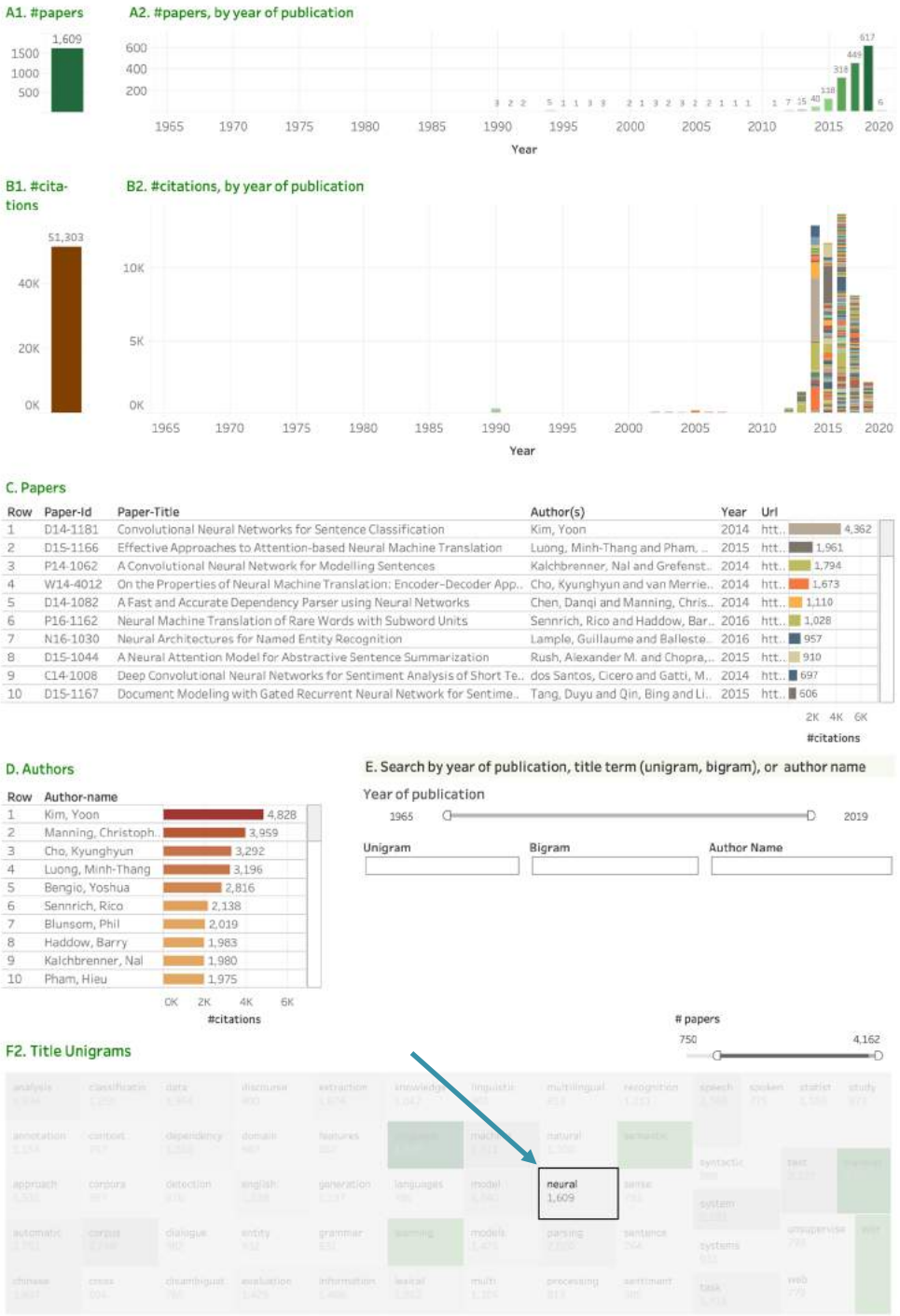
NLP Scholar Dashboard

Clicking on ‘statistical’ in the
Title Unigrams treemap (F2)



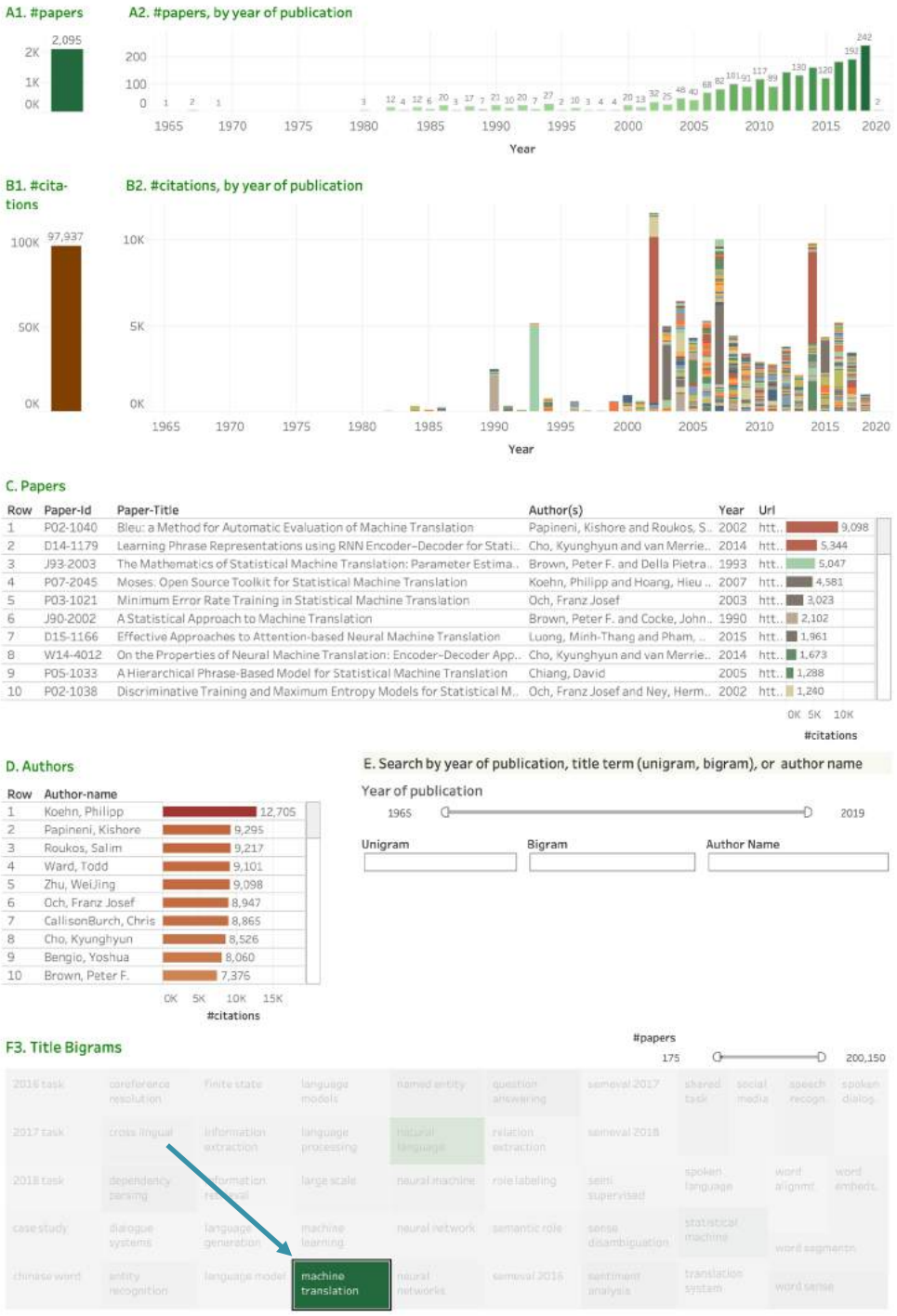
NLP Scholar Dashboard

Clicking on ‘neural’ in the
Title Unigrams treemap (F2)



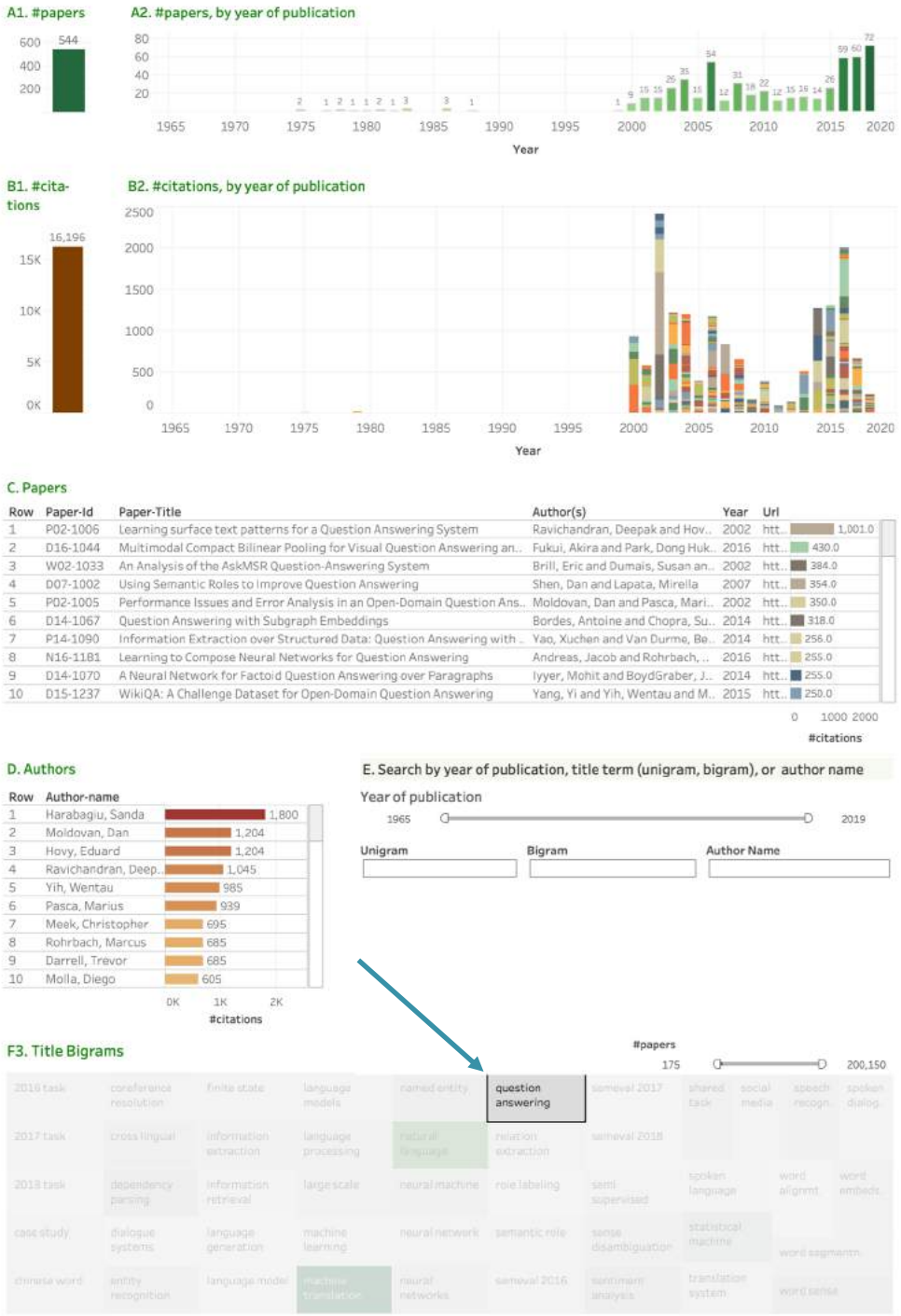
NLP Scholar Dashboard

Clicking on `machine translation' in the Title Bigrams treemap (F3)



NLP Scholar Dashboard

Clicking on ‘question answering’ in the Title Bigrams treemap (F3)

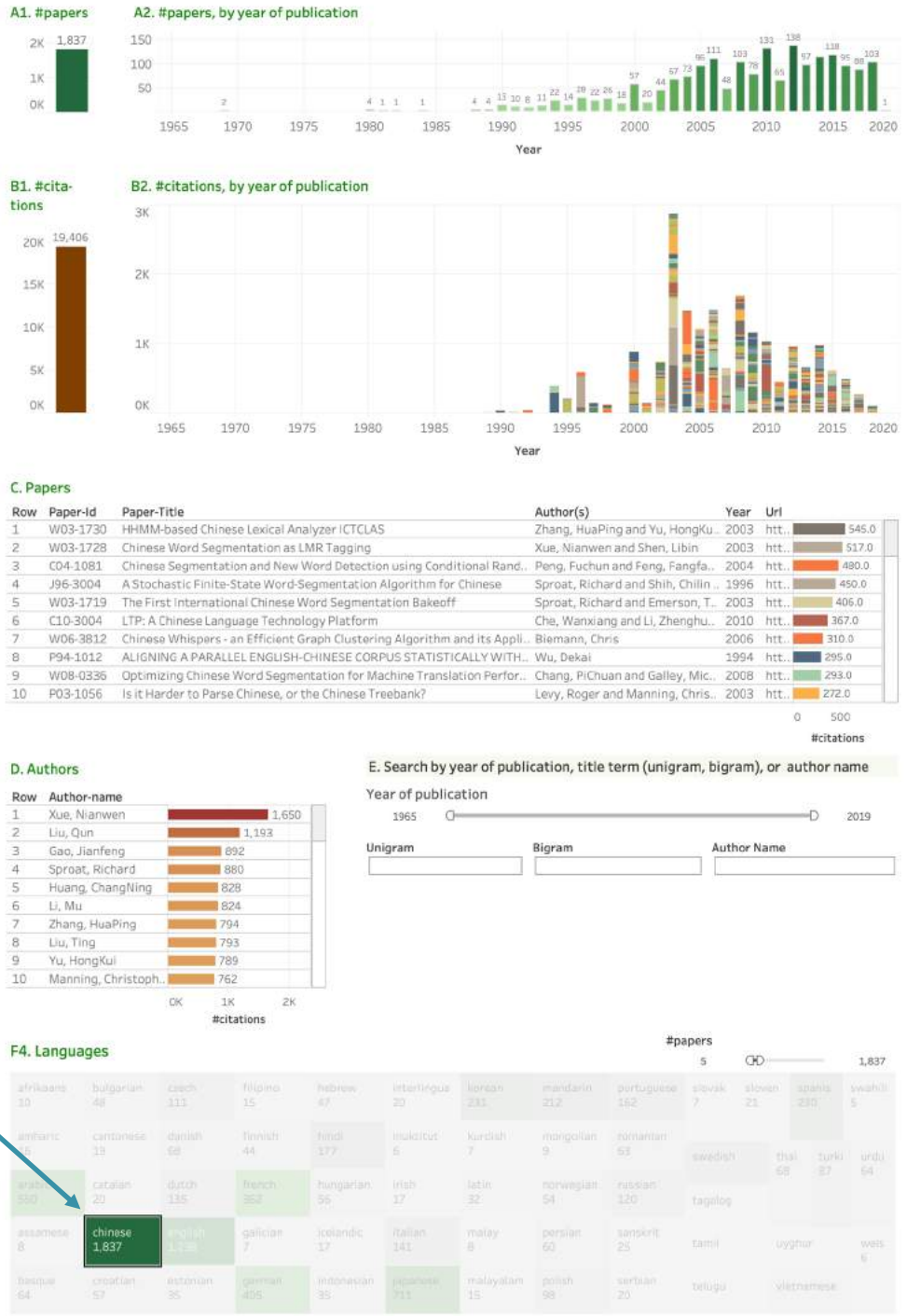


NLP Scholar Dashboard

Clicking on ‘word embeddings’ in the Title Bigrams treemap (F3)

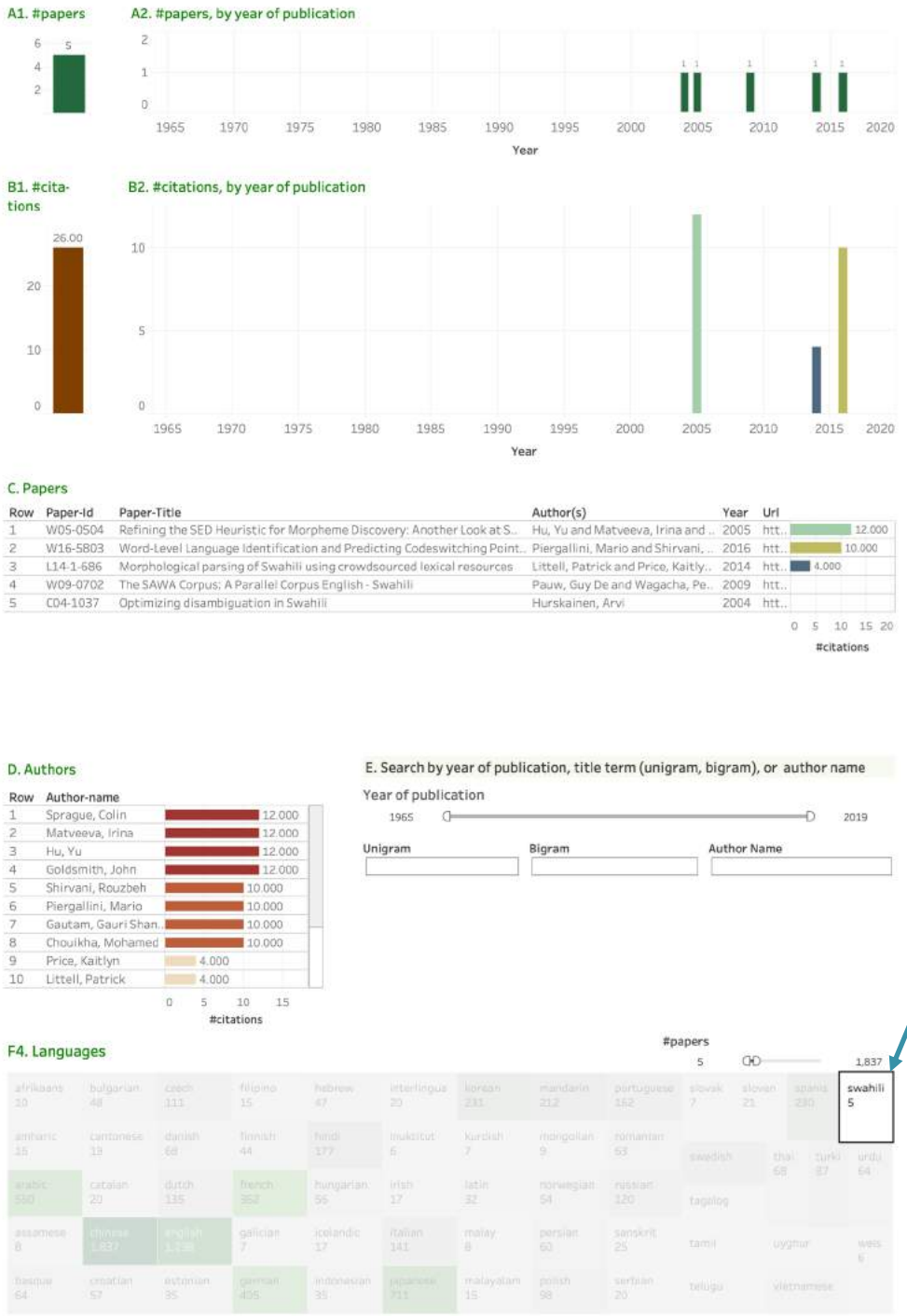
NLP Scholar Dashboard

Clicking on 'Chinese' in the
Languages treemap (F4)



NLP Scholar Dashboard

Clicking on 'Swahili' in the
Languages treemap (F4)



Summary

NLP Scholar—an interactive visual explorer for the ACL Anthology

- access to citation information from Google Scholar
- interconnected interactive visualizations (multiple dashboards)
- quickly and efficiently search for relevant related work

Created by Symbolon from Noun Project

Future Work:

- automatically identifying related papers
- analyze NLP papers that are published outside of the ACL Anthology
- compare patterns of citations in NLP with those in other fields
- develop richer ways of capturing scholarly impact

Project page for NLP Scholar: http://saifmohammad.com/WebPages/nlpscholar.html

- data
- Interactive visualizations
- limitations and ethical considerations

Contact: ✉ Saif.Mohammad@nrc-cnrc.gc.ca 🐦 @SaifMMohammad