Data Cleaning and Mapping for Bibliometric Studies

Zehra Taşkın & Güleda Doğan

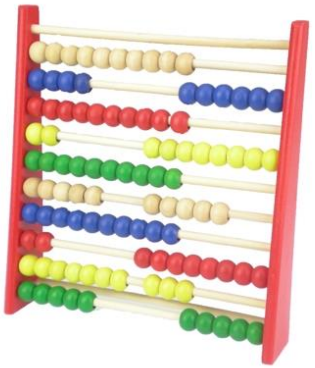
Bibliometrics?

- Citation; relationship between cited and citing documents
- It can be used to evaluate relevant sources, promote authors, provide a more extensive reading list

Bibliometrics?

- Set of methods to analyze literature quantitatively.
- 1873 – Shepherd's Citations – first indexing of citations
- 1969 – Alan Pritchard (Statistical Bibliography or Bibliometrics)
- «the application of mathematics and statistical methods to books and other media of communication»
- Citation analyses are commonly used bibliometric method.

Counting Citations for Bibliometrics



- Is it the best way to evaluate science?
  - What about negative citations!
    - I can write a paper with many mistakes, how can you evauate my citations? There can be too many criticism on my paper's quality.
    - Are citation counts made my study valuable?
  - What about self-citations?
  - What about methodological papers?

  - **And what about data quality?**

Data Sources

- Mostly used and well known citation sources are;
  - Web of Science
  - Scopus

- You can download the data from anywhere you want.
  - Web
  - Other databases
- You can create your own dataset

THE ORIGINS OF DATA PROBLEMS

The Origins of Data Problems

- Natural language indexing
  - There is a hospital in Turkey named «Dr. Sami Ulus»
  - An author from this hospital wrote his affiliation as «Dr. Saaaami Ulus»

Addresses:

[1] Hacettepe Univ, Fac Med, Dept Paediat, Paediat Haemat

[2] Dr Saaaaami Ulus Childrens Hosp, Pediat Hematol Sec

[3] Hacettepe Univ, Fac Med, Dept Pediat, Pediat Hematol S

The Origins of Data Problems

- Errors occured during digitisation processes
  - The university's name; «Fırat Univ»
  - It can be seemed on Web of Science as «F1rat Univ»

Addresses:
[1] Ankara Univ, Dept Pharmacognosy, Fac Pharm, TR-06100 Tandogan, Turkey
[2] F1rat Univ, Fac Sci & Educ, TR-23119 Elazig, Turkey

The Origins of Data Problems

- Existence of the authors with the same name
- Various abbreviations for the names
  - a single name can represent different authors and an individual author can publish papers under multiple names

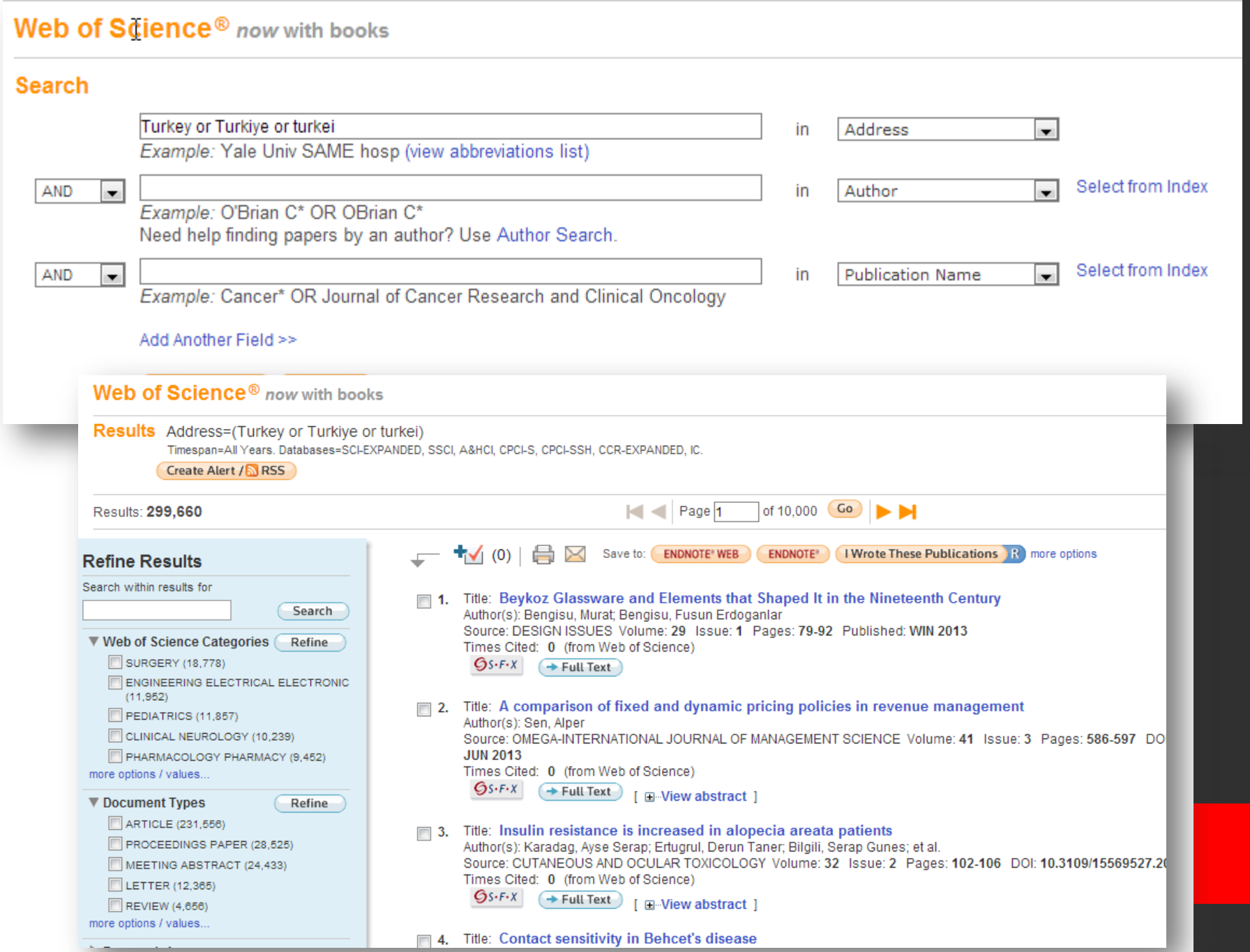
The Origins of Data Problems

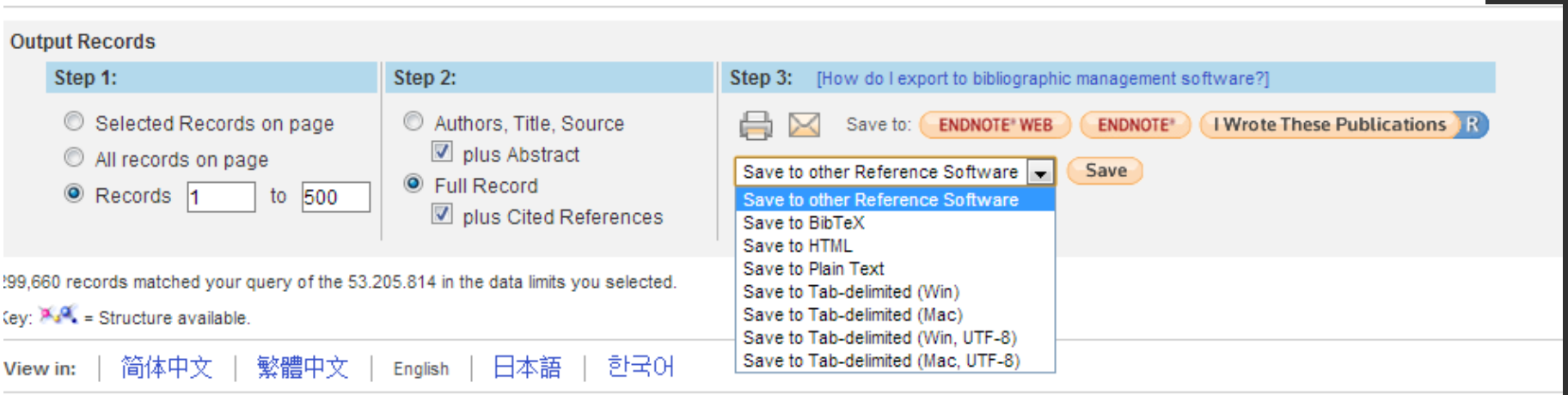
- Other indexing mistakes
  - Indexers can not know everything with the details
  - It is inevitable not to make mistakes with the natural language indexing.

Effects of Unreliable Bibliometric Data

University	Pub. Count	Errors (N)	Errors (%)
Hacettepe University	19,166	340	1.8
İstanbul University	16,390	1,691	10.3
Ankara University	13,275	224	1.7
METU	11,201	102	0.9
Ege University	9,428	654	6.9
Gazi University	9,281	85	0.9
İstanbul Technical Univ.	8,613	215	2.5
Dokuz Eylül University	6,069	210	3.5
Atatürk University	5,816	46	0.8
GATA	5,300	639	12.1

GATHERING DATA


Web of Science® now with books
Search
Turkey or Turkiye or turkei
in
Address
Example: Yale Univ SAME hosp (view abbreviations list)
AND
in
Author
Select from Index
Example: O'Brian C* OR OBrian C*
Need help finding papers by an author? Use Author Search.
AND
in
Publication Name
Select from Index
Example: Cancer* OR Journal of Cancer Research and Clinical Oncology
Add Another Field >>
Web of Science® now with books
Results Address=(Turkey or Turkiye or turkei)
Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, CCR-EXPANDED, IC.
Create Alert / RSS
Results: 299,660
Page 1 of 10,000 Go
Refine Results
Search within results for
Search
Web of Science Categories Refine
SURGERY (18,778)
ENGINEERING ELECTRICAL ELECTRONIC (11,952)
PEDIATRICS (11,857)
CLINICAL NEUROLOGY (10,239)
PHARMACOLOGY PHARMACY (9,452)
more options / values...
Document Types Refine
ARTICLE (231,556)
PROCEEDINGS PAPER (28,525)
MEETING ABSTRACT (24,433)
LETTER (12,365)
REVIEW (4,656)
more options / values...
(0)
Save to: ENDNOTE® WEB ENDNOTE® I Wrote These Publications more options
1. Title: Beykoz Glassware and Elements that Shaped It in the Nineteenth Century
Author(s): Bengisu, Murat; Bengisu, Fusun Erdoganlar
Source: DESIGN ISSUES Volume: 29 Issue: 1 Pages: 79-92 Published: WIN 2013
Times Cited: 0 (from Web of Science)
Full Text
2. Title: A comparison of fixed and dynamic pricing policies in revenue management
Author(s): Sen, Alper
Source: OMEGA-INTERNATIONAL JOURNAL OF MANAGEMENT SCIENCE Volume: 41 Issue: 3 Pages: 586-597 DO
JUN 2013
Times Cited: 0 (from Web of Science)
Full Text [View abstract]
3. Title: Insulin resistance is increased in alopecia areata patients
Author(s): Karadag, Ayse Serap; Ertugrul, Derun Taner; Bilgili, Serap Gunes; et al.
Source: CUTANEOUS AND OCULAR TOXICOLOGY Volume: 32 Issue: 2 Pages: 102-106 DOI: 10.3109/15569527.2
Times Cited: 0 (from Web of Science)
Full Text [View abstract]
4. Title: Contact sensitivity in Behcet's disease


Output Records
Step 1:
Selected Records on page
All records on page
Records 1 to 500
Step 2:
Authors, Title, Source
plus Abstract
Full Record
plus Cited References
Step 3: [How do I export to bibliographic management software?]
Save to: ENDNOTE® WEB ENDNOTE® I Wrote These Publications R
Save to other Reference Software
Save
Save to other Reference Software
Save to BibTeX
Save to HTML
Save to Plain Text
Save to Tab-delimited (Win)
Save to Tab-delimited (Mac)
Save to Tab-delimited (Win, UTF-8)
Save to Tab-delimited (Mac, UTF-8)
99,660 records matched your query of the 53.205.814 in the data limits you selected.
Key: = Structure available.
View in: 简体中文 繁體中文 English 日本語 한국어

	A	B	C	D	E	F
1	PT	AU	BA	BE	GP	AF
2	J	Bengisu, M; Bengisu, FE				Bengisu, Murat; Bengisu, Fusun Erdoganlar
3	J	Sen, A				Sen, Alper
4	J	Karadag, AS; Ertugrul, DT; Bilgili, SG; Takci, Z; Tutal, E; Yilmaz, H				Karadag, Ayse Serap; Ertugrul, Derun Taner; Bilgili,
5	J	Demirsoy, EO; Kiran, R; Ozturk, B; Akturk, AS; Etiler, N				Demirsoy, Evren Odyakmaz; Kiran, Rebiay; Ozturk,
6	J	Gunay, U; Gunduz, K; Ermertcan, AT; Kandiloglu, AR				Gunay, Umran; Gunduz, Kamer; Ermertcan, Aylin Tu
7	J	Onder, HI; Turan, H; Kilic, AC; Kaya, M; Tunc, M				Onder, Halil Ibrahim; Turan, Hakan; Kilic, Ali Cagri;
8	J	Gunduz, K; Coban, M; Ozturk, F; Ermertcan, AT				Gunduz, Kamer; Coban, Mutlu; Ozturk, Ferdi; Erme
9	J	Karadag, AS; Bilgili, SG; Calka, O; Onder, S; Kosem, M; Burakgazi-Dalkilic, E				Karadag, Ayse Serap; Bilgili, Serap Gunes; Calka, Or
10	J	Tecimer, RS; Yildiz, KD; Akturk, AS; Bilen, N				Tecimer, Rukiye Selin; Yildiz, Kursat Demir; Akturk,
11	J	Turan, H; Okur, M; Kaya, E; Gun, E; Aliagaoglu, C				Turan, Hakan; Okur, Mesut; Kaya, Ertugrul; Gun, Em
12	J	Altun, ML; Yilmaz, BS; Orhan, IE; Citoglu, GS				Altun, M. Levent; Yilmaz, Betul Sever; Orhan, Ilkay
13	J	Berhow, MA; Polat, U; Glinski, JA; Glensk, M; Vaughn, SF; Isbell, T; Ayala-Diaz, I; Marek, L; Gardner, C				Berhow, Mark A.; Polat, Umit; Glinski, Jan A.; Glens
14	J	Hatipoglu, G; Sokmen, M; Bektas, E; Daferera, D; Sokmen, A; Demir, E; Sahin, H				Hatipoglu, Gonul; Sokmen, Munevver; Bektas, Ersa
15	J	Ayrilmis, N; Kaymakci, A				Ayrilmis, Nadir; Kaymakci, Alperen
16	J	Aytekin, O; Zongur, U; Halici, U				Aytekin, O.; Zongur, U.; Halici, U.
17	J	Yilmaz, MT; Karaman, S; Kayacier, A				Yilmaz, Mustafa Tahsin; Karaman, Safa; Kayacier, Ah
18	J	Cakirli, O; Ibanoglu, C; Sipahi, E				Cakirli, Omur; Ibanoglu, Cafer; Sipahi, E.
19	J	Yilmaz, M; Selam, SO; Sato, B; Izumiura, H; Bikmaev, I; Ando, H; Kambe, E; Keskin, V				Yilmaz, M.; Selam, S. O.; Sato, B.; Izumiura, H.; Bikm
20	J	Yaman, U; Dolen, M				Yaman, U.; Dolen, M.
21	J	Ozkan, C; Karaesmen, F; Ozekici, S				Ozkan, Can; Karaesmen, Fikri; Ozekici, Suleyman
22	J	Zafer, A				Zafer, Agacik
23	J	Muter, I; Birbil, SI; Bulbul, K; Sahin, G; Yenigun, H; Tas, D; Tuzun, D				Muter, Ibrahim; Birbil, S. Ilker; Bulbul, Kerem; Sahi

C1	RP
[Ihtiyar, Enver; Erkasap, Serdar; Karakas, Baris R.; Yasar, Fatih N.] Eskisehir Osmangazi Univ, Sch Med, Dept Gen Surg, TR-26480 Eskisehir, Turkey; [Pasaoglu, Ozgul] Eskisehir Osmangazi Univ, Sch Med, Dept Pathol, TR-26480 Eskisehir, Turkey	Ihtiyar, E (reprint author), Eskisehir O
[Kesemenli, Cumhur Cevdet; Memisoglu, Kaya] Kocaeli Univ Sch Med, Dept Orthopaed & Traumatol, Izmit, Kocaeli, Turkey; [Necmioglu, Serdar] Dicle Univ Sch Med, Dept Orthopaed & Traumatol, Diyarbakir, Turkey; [Kayikci, Cuma] Guneydogu Hosp, Orthopaed Clin, Diyarbakir, Turkey	Memisoglu, K (reprint author), Kocae
[Tari, Rabia] Adnan Adivar Cad Haseki Egitim Arastirma Hastanes, Cerrahisi Klin Fatih, Istanbul, Turkey; [Ulusal, Ismail; Tari, Rabia; Ozturk, Gulsah; Aycicek, Ezgi; Bilge, Turgay] Haseki Training & Res Hosp, Dept Neurosurg, Istanbul, Turkey; [Aktar, Fadime] Istanbul Univ, Dept Histol & Embryol, Istanbul Fac Med, Istanbul, Turkey; [Kotil, Kadir] Istanbul Training & Res Hosp, Dept Neurosurg, Istanbul, Turkey; [Kiris, Talat] Istanbul Univ, Dept Neurosurg, Istanbul Fac Med, Istanbul, Turkey	Tari, R (reprint author), Adnan Adivar
[Kanat, Ayhan] Rize Univ, Dept Neurosurg, Fac Med, Rize, Turkey; [Yilmaz, Adem; Musluman, Murat] Sisli Res & Educ Hosp, Dept Neurosurg, Istanbul, Turkey; [Aydin, Mehmet D.; Altas, Sare; Gursan, Nesrin] Ataturk Univ, Fac Med, Dept Neurosurg, Erzurum, Turkey; [Altas, Sare; Gursan, Nesrin] Ataturk Univ, Fac Med, Dept Pathol, Erzurum, Turkey;	

CLEANING DATA

Cleaning Data

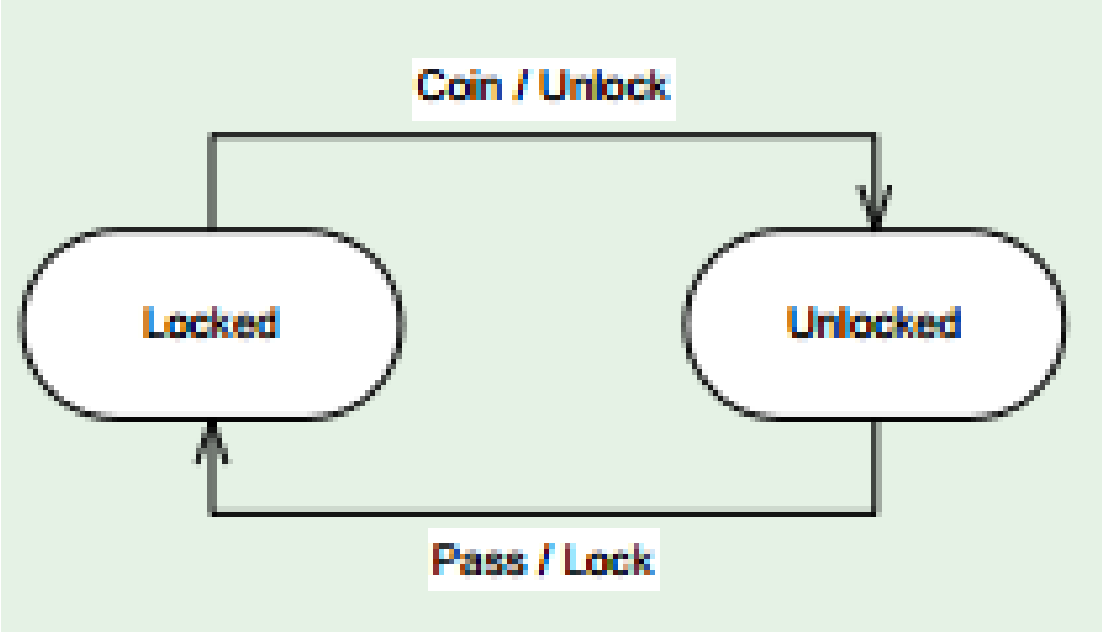
- Cleaning data by manually is impractical and time consuming
- Cleaning 200,000 data can occupy at least 6 month
- Is it worth it?

Cleaning Data

- There is different techniques in the literature to determine mistakes and non-stardardized usages of the things (author, organization or journal name) and standardize the entities automatically.
- Well known techniques are;
  - Finite State Technique
  - Clustering Technique

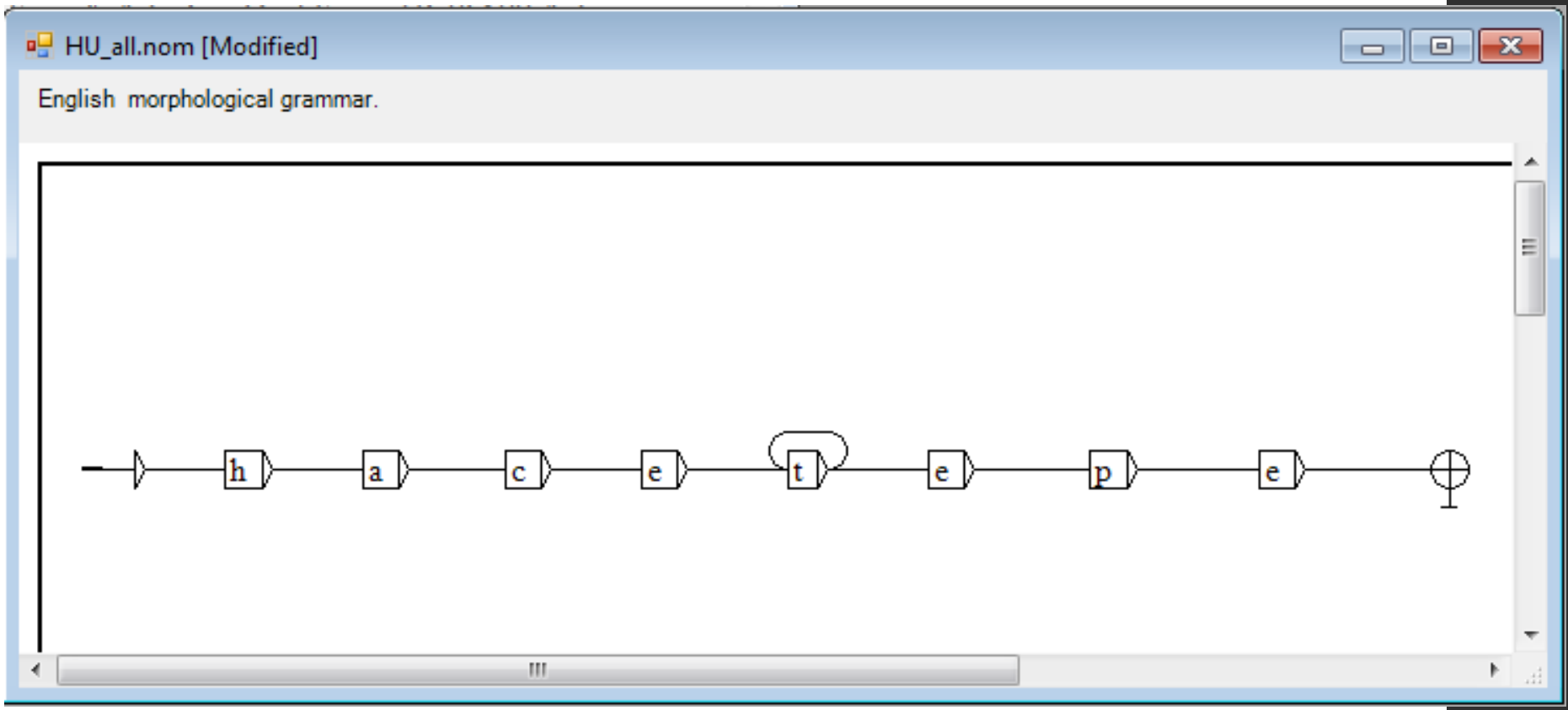
Finite State Technique

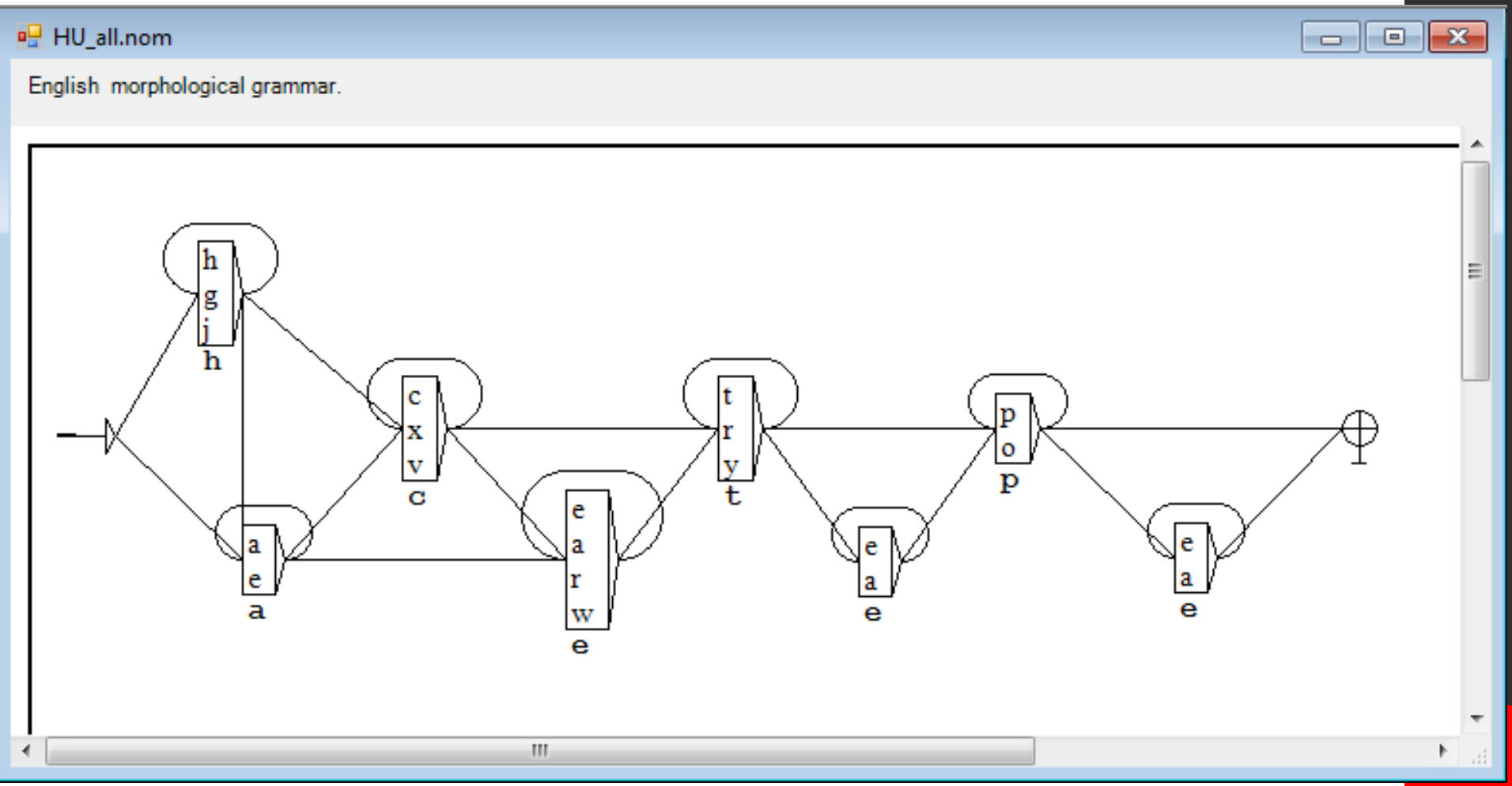
- It is based on the sets of states and their relationships with each other
- It is like metro turnstile;



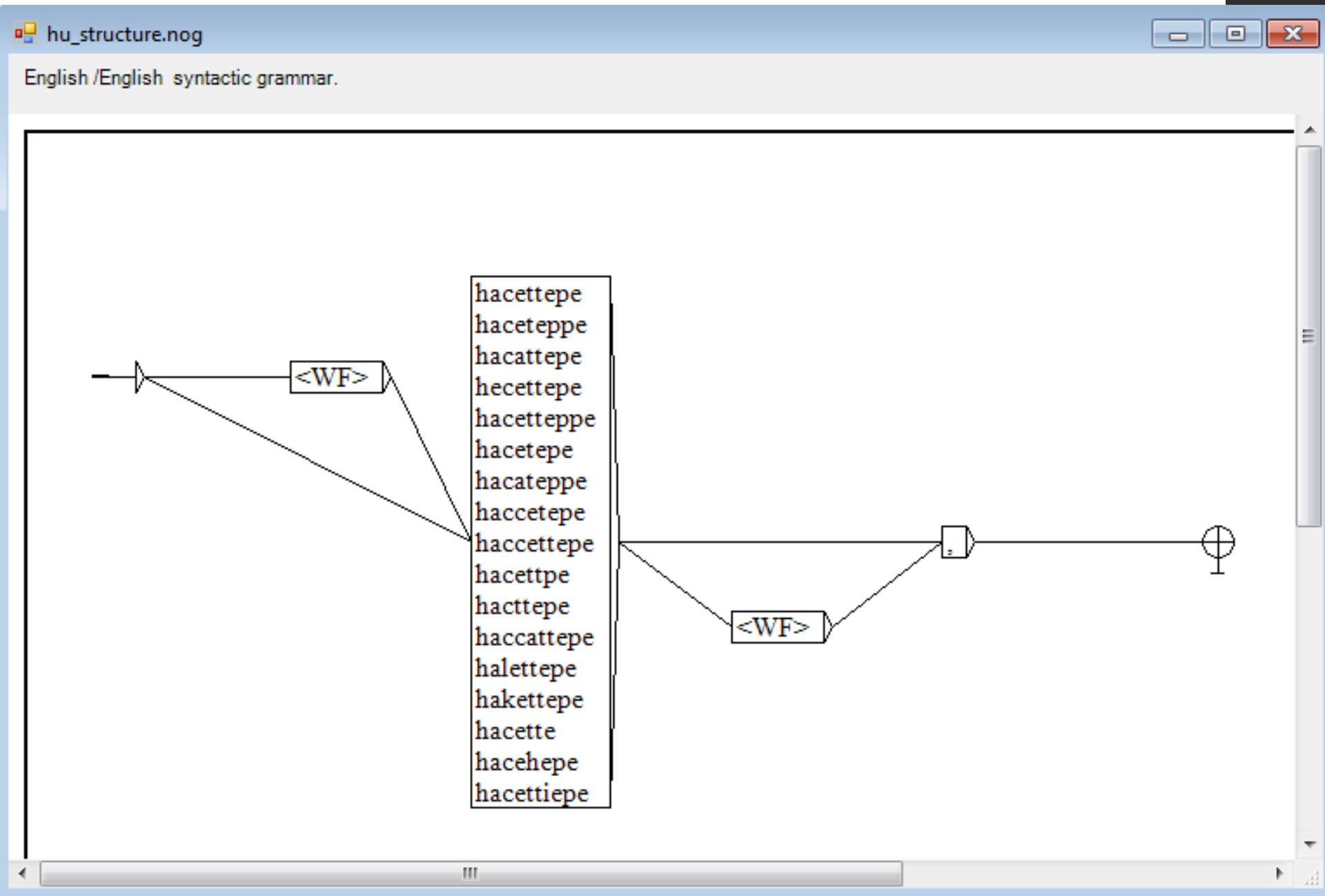
Finite State Technique

- Finite state technique can be implemented to large amount of data by finite state transducers easily.
- Nooj transducer (http://www.nooj4nlp.net)


HU_all.nom [Modified]
English morphological grammar.
h
a
c
e
t
e
p
e

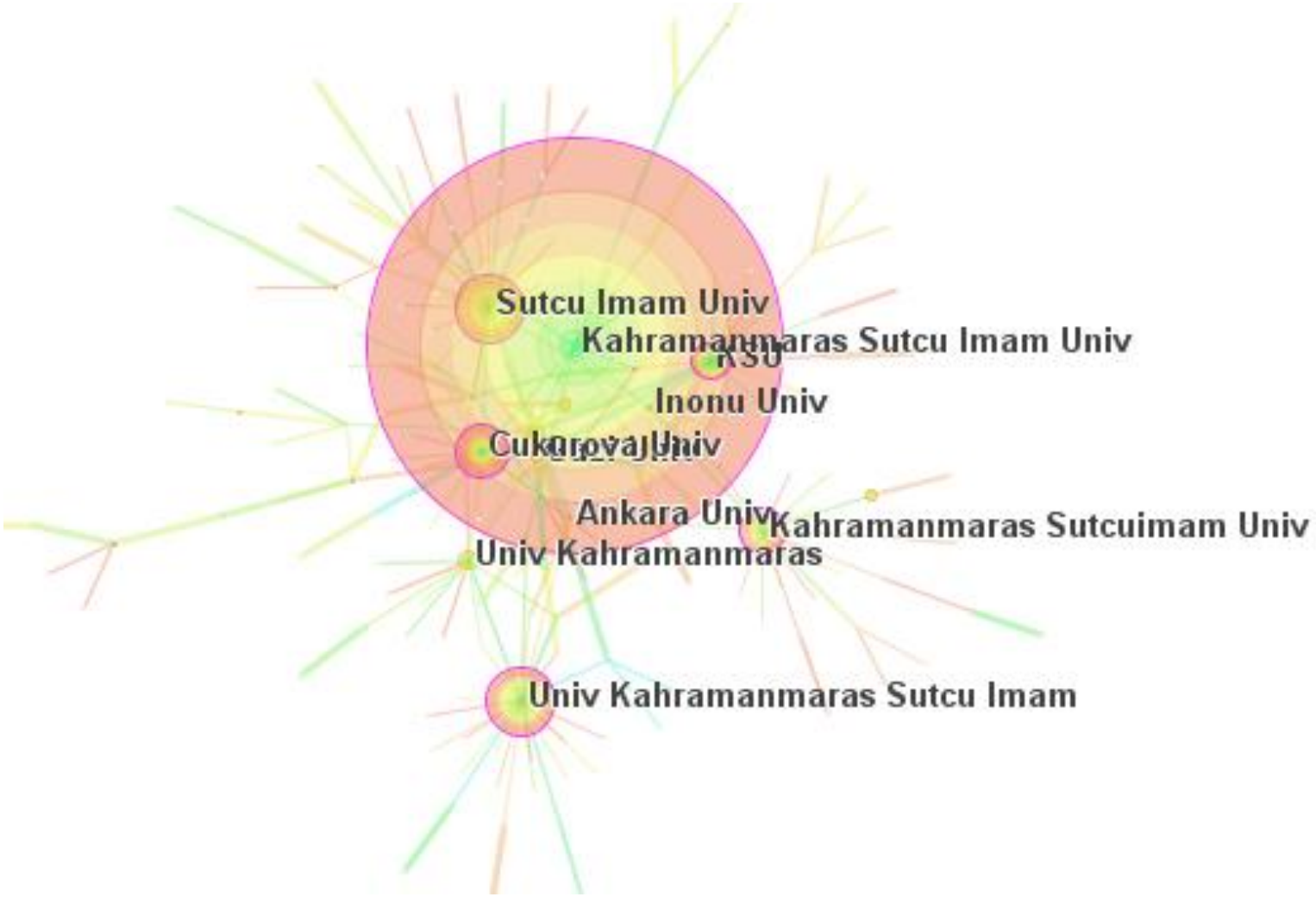

HU_all.nom
English morphological grammar.
h
g
j
h
c
x
v
c
t
r
y
t
p
o
p
a
e
a
e
a
r
w
e
e
a
e
e
a
e

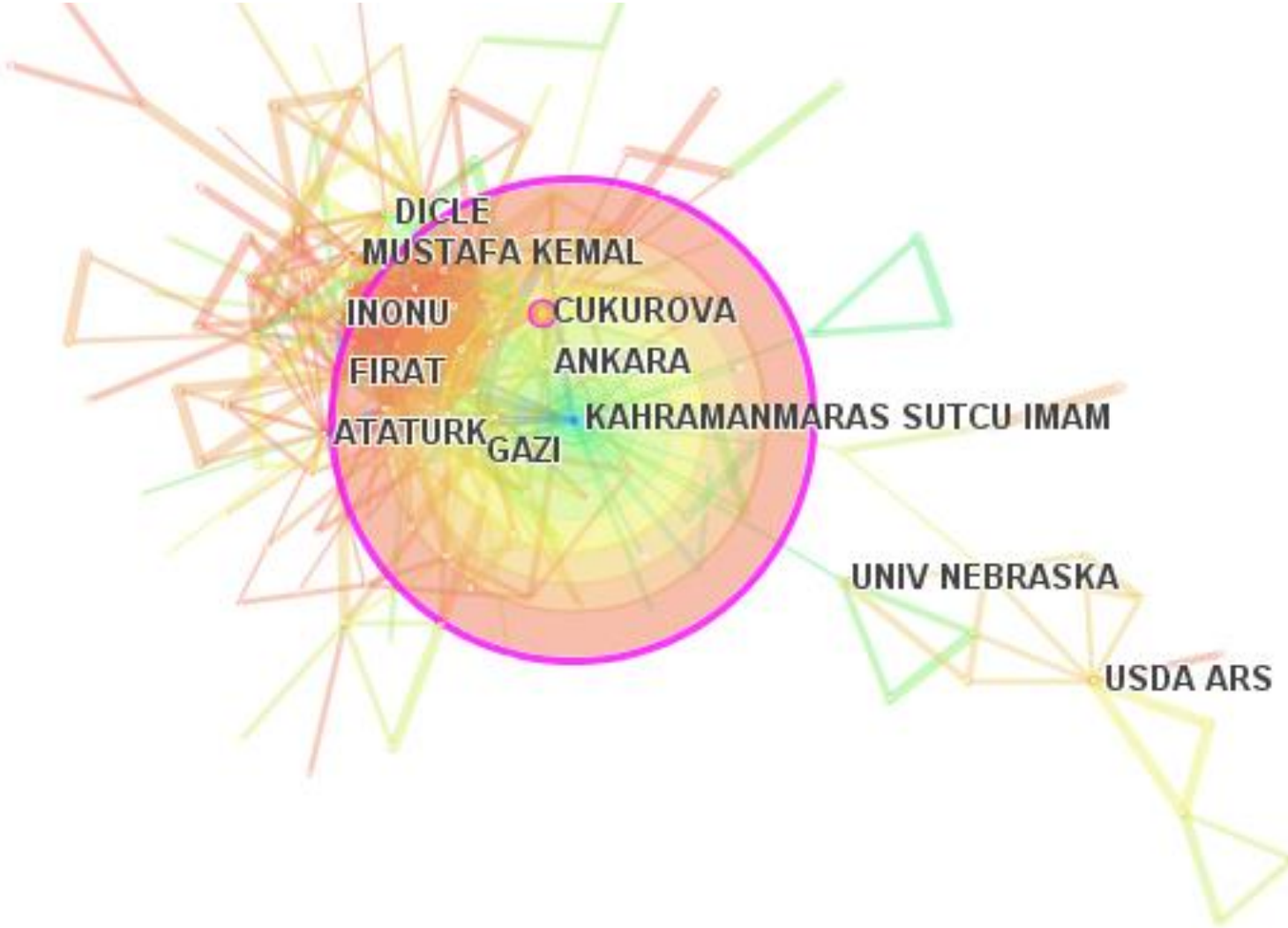
Term	Freq	Term	Freq
Hacettepe	21.024	Hacateppe	2
HACETTEPE	4.658	HACCETEPE	2
HACETEPPE	9	HACCETTEPE	2
Hacattepe	7	HACETTPE	2
HECETTEPE	5	Hacttepe	1
HACETTEPPE	4	Haccattepe	1
Hacetepe	3	Haccettepe	1
Hecettepe	3	HACETEPE	1


hu_structure.nog
English /English syntactic grammar.
<WF>
hacettepe
haceteppe
hacattepe
hecettepe
hacetteppe
hacetepe
hacateppe
haccetepe
haccettepe
hacettpe
hacttepe
haccattepe
halettepe
haketttepe
hacette
hacehepe
hacettiepe
<WF>

Hacettepe Univ	Hacettepe Adult Hosp	Hacettepe Tip Fak
Univ Hacettepe	Hacettepe Child Hosp	Hacattepe Univ
Hacettepe Childrens Hosp	Hacettepe Cocuk Hastabanesi	Hacettepe Children Hosp
Hacettepe Med Sch	Hacttepe Univ	Hecettepe Univ
Hacettepe U	Hakettepe Childrens Hosp	Hacetepe Univ
Hacettepe Sch Med	Halettepe Univ	Hacetteppe Univ
Hacettepe Med Fac	Univ Haceteppe	Hacettpe Univ
Hacettepe Fac Med	Hacettepe Cocuk Hastahanesi	Hosp Hacettepe
Hacettepe Hastaneleri	Hacettepe Cocuk Hastanesi	Hacattepe Univ Hosp
Hacettepe Hosp	Hacettepe Cocuk Hastenesi	Haccattepe Univ
Hacettepe Hastanesi	Hacettepe Eriskin Hastanesi	Haccetepe Fac Med
Hacettepe Oncol Inst	Hecettepe Childrens Hosp	Haccettepe Univ Hosp
Haceteppe Univ	Klinikum Hacettepe	Hacettepe Med Acad
Hecettepe Univ	Hacettepe Inst Oncol	Hacettepe Kuniv Hastaneleri
Hacettepe Med Ctr	Unit Hacettepe	Hacehepe Univ
Hacette Unıv Sociales Hacettepe	Hacettepe Technopolis	Haceteppe Childrens Hosp

MAPPING FOR BIBLIOMETRIC STUDIES


Sutcu Imam Univ
Kahramanmaras Sutcu Imam Univ
KSU
Inonu Univ
Cukurova Univ
Ankara Univ
Kahramanmaras Sutcuimam Univ
Univ Kahramanmaras
Univ Kahramanmaras Sutcu Imam


DICLE
MUSTAFA KEMAL
INONU
CUKUROVA
FIRAT
ANKARA
KAHRAMANMARAS SUTCU IMAM
ATATURK
GAZI
UNIV NEBRASKA
USDA ARS