

סילבוס - תוכנית הוראה לקורס  
שם הקורס:

## שער אל מדעי הרוח הדיגיטליים

שם המרצה: ד"ר אבי שמידמן

שם המחלקה: ספרות עם ישראל

מספר הקורס: 13-6190-01

שם הקורס באנגלית:

## Digital Humanities: A Primer

סוג הקורס: הרצאה

היקף נ"ז: 2

שנת לימודים: תשפ"ה

סמסטר: א

יום ושעה: רביעי 18:00

שעת קבלה: בתיאום מראש

מייל מרצה: [avi.shmidman@biu.ac.il](mailto:avi.shmidman@biu.ac.il)

קישור לאתר למדה: <https://lemida.biu.ac.il/>



## תיאור הקורס ומטרות למידה

### תקציר הקורס

הקורס נועד להנגיש את העולם של מדעי הרוח הדיגיטליים לסטודנטים שאין להם בהכרח שום רקע במדעי המחשב או בתכנות מחשבים.

### מטרות/תוצרי הלמידה:

#### א. ידע

1. הלומדים ילמדו מה יש לעולם של "מדעי הרוח הדיגיטליים" להציע להם בתור חוקרי מדעי הרוח, מעבר למאגרי מידע גרידא
2. הלומדים ילמדו מה יש בכוחו של מחשב להכריע, ולעומת זאת, מתי אין לסמוך על הכרעות של המחשב.

#### ב. מיומנויות

1. לברר ענייני ייחוס של טקסטים אנונימיים או של טקסטים שמחברם שנוי במחלוקת
3. לעמוד על המאפיינים הסגנוניים הייחודיים של קורפוס טקטואלי כלשהו ("סטיילומטריה")
4. להפיג את העמימות של טקסט עברי לא-מנוקד (למשל, להבחין באופן אוטומטי בין "אָת" ל"אַת" ובין "אָם" ל"אַם")
5. לנצל כלים טכנולוגיים כדי לעמוד על ההבדלים בין הדפסות שונות של יצירת ספרות נתונה
6. לבנות מערכות OCR מתקדמות שמותאמות באופן ייעודי לקורפוס מסוים
7. לנצל רשתות עצביות (רשתות ניורוניות) כדי לאתר תופעות ספרותיות ספיציפיות בקורפוסים של "big data"
8. לאתר מקבילות חלקיות בין קורפוסים טקטואליים שונים.



## למידה פעילה - תכנון מהלך השיעורים:

מס' השיעור	נושא השיעור	למידה פעילה	קריאה/ צפיה נדרשת	הערכה תהליכית/מעצבת
1	Big Data vs. Smart Data		Schoch, Data	
2-3	הפגת העמימות של הטקסט העברי: האתגר ופתרונו. ניקוד אוטומטי, תיוג מורפולוגי, תיוג תחבירי, הבחנה בין משמעים; פתיחה אוטומטית של ראשי תיבות.		Karpathy, RNN; Shmidman et al., Nakdan; Gershuni and Pinter, Hebrew Diacritics; Tsarfaty et al., Hebrew NLP; Tsarfaty, Integrated; Rubin, Abbreviations; HaCohen-Kerner et al., Haads; Rubinstein and Shmidman, Chronolect	
4	שיכוני מילים (word embeddings)		Widdows, Geometry and Meaning; Mikolov et al., Efficient; Devlin et al., BERT; Seker et al., AlephBERT; Shmidman et al., Challenge Sets; Gonen et al., Word2vec; Liebeskind et al., Construction	
5-6	דיגיטציה של טקסטים: סריקה ופיענוח (OCR).		Mahpod and Keller, Rashi Scripts; Bulacu-Schomaker, Automatic	
7	זיהוי אוטומטי של מקבילות בין טקסטים		Shmidman, Biblical Citations; Shmidman et al., Identification; Bar-Asher Siegal and Shmidman, Reconstruction; Schorr et al., ViS-Á-ViS	
8	נוסחים שונים של טקסט אחד – אלגוריתמים של sequence alignment ויישומיהם במדעי הרוח הדיגיטליים		Och-Ney, Systematic; Brill et al., FAST	תרגיל א: הערכה של סינופסיס אוטומטי
9	מהדורות מדעיות דיגיטליות		מרינגר מיליקובסקי, מילים שקולות; Ide-Véronis, TEI; Deegan-Sutherland, Text Editing; Pierazzo, Digital Editing	

10-11	סטיילומטריה: זיהוי המרכיבים הסגנוניים הייחודיים בטקסט נתון	מינץ-מנור ומרינברג- מיליקובסקי, מחקר חישובי Moretti, Graphs; Moretti, Distant; Netzer et al., Evaluating; Munz-Manor, Computational Study	תרגיל ב: סטיילומטריה
12-13	בירור ייחוס של טקסטים	Koppel-Schler, Stylistic Idiosyncrasies; Koppel et al., Unmasking; Kestemont, Function Words; Fuchs, Moed Katan	
14	זיהוי אוטומטי של שגיאות במאגרי מידע טקסטואליים	Mirkin, Agnon; Al-Azawi, Statistical Language; Suissa, Error Correction	

- ייתכנו שינויים בסילבוס בהתאם לקצב ההתקדמות ואפקטיביות הלמידה



## ציון סופי

תיאור התוצר	משקל בציון הסופי
תרגיל א	5% מהציון הסופי
תרגיל ב	5% מהציון הסופי
מבחן סופי	90% מהציון הסופי



## דרישות הקורס

שם הקורס	מס' הקורס

## ביבליוגרפיה: תכנים לקריאה, צפיה והאזנה

מינץ-מנור, א' וא' מרינברג-מיליקובסקי (עורכים), מחקר חישובי במדעי הרוח: אסופת מאמרים, רעננה תשפ"ג

מרינברג-מיליקובסקי, א', מילים שקולות: צעדים ראשונים במחקר הספרות החישובי, רעננה תשפ"ג

- Adler, Morphology = Meni Adler, "Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach," PhD thesis, Ben-Gurion University, 2007
- Al-Azawi, Statistical Language = Mayce Al-Azawi, "Statistical Language Modeling for Historical Documents using Weighted Finite-State Transducers and Long Short-Term Memory," Phd Thesis, Kaiserslautern, 2015
- Al-Haj, Multiword Expressions = Hassan Al-Haj, "Hebrew Multiword Expressions: Linguistic Properties, Lexical Representation, Morphological Processing, and Automatic Acquisition," PhD Thesis, Haifa 2009
- Bar-Asher Siegal and Shmidman, Reconstruction = Michal Bar-Asher Siegal and Avi Shmidman, "Reconstruction of the Mekhilta Deuteronomy Using Philological and Computational Tools," *Journal of Ancient Judaism* 9, 1 (2019), 2-25
- Belinkov-Glass, Arabic Diacritization = Y. Belinkov and J. R. Glass, "Arabic Diacritization with Recurrent Neural Networks," *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015, 2281-2285.
- Belinkov et al., Historical = Yonatan Belinkov, Alexander Magidow, Maxim Romanov, Avi Shmidman and Moshe Koppel, "A Large-Scale Historical Arabic Corpus," *Proceedings of the COLING Workshop on Language Technology and Tools for Digital Humanities*, December 2016
- Brill et al., FAST = Oran Brill, Moshe Koppel and Avi Shmidman, "FAST: Fast and Accurate Synoptic Texts", *Digital Scholarship in the Humanities* (forthcoming).
- Bulacu-Schomaker, Automatic = M. Bulacu, and L. Schomaker, "Automatic handwriting identification on medieval documents," *Int. Conf. on Image Analysis and Processing*, 2007
- Choueka, Needles = Yaacov Choueka, "Looking for Needles in a Haystack or: Locating Interesting Expressions in Large Textual Databases," *Proceedings of the International Conference on User-Oriented Content-Based Text and Image Handling (RIAO)*, Cambridge, Mass. (1988), 609-623

- Choueka, Responsa = Yaacov Choueka, “Responsa – A Full-Text System with Linguistic Components for Large Corpora,” in: Quemada, B. and Zampolli, A. (eds.), *Computational Lexicology and Lexicography*, Giardini Editions, Pisa, 1990, 181-217
- Choueka et al., Collocations = Y. Choueka, S. T. Klein, E. Neuwitz, “Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus,” *ALLC J.* 4 (1983), 34-38
- Cummings, TEI = James Cummings, “The Text Encoding Initiative and the study of literature,” in: *A Companion to Digital Literary Studies*, ed. Ray Siemens and Susan Schreibman, Oxford: Blackwell, 2013, 451-476
- Deegan-Sutherland, Text Editing = Marilyn Deegan and Kathryn Sutherland (eds.), *Text Editing, Print, and the Digital World*, Surrey, England and Burlington, VT: Ashgate Publishing, 2009
- Devlin et al., BERT = Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, 4171-4186, 2019.
- Fuchs, Moed Katan = Yacov Fuchs, “Rashi’s Commentary to Tractate Moed Katan: Determining Authorship and Methods of Transmission and Formation”, Ph.D. Thesis, Bar-Ilan University [Hebrew], 2007
- Gershuni and Pinter, Hebrew Diacritics = Elazar Gershuni and Yuval Pinter, “Restoring Hebrew Diacritics Without a Dictionary”, arXiv:2105.05209
- Goldberg, Neural Networks = Yoav Goldberg, *Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies*. San Rafael, CA, 2017.
- Gonen et al., Word2vec = Hila Gonen, Ganesh Jawahar, Djamé Seddah, Yoav Goldberg, “Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- Goodwin-Holbo, Response = Jonathan Goodwin and John Holbo (ed.), *Reading Graphs, Maps, and Trees: Responses to Franco Moretti*, Glassbead Books 2011.
- HaCohen-Kerner et al., Haads = Y. HaCohen-Kerner, and A. Kass, and A. Peretz, “Haads: A Hebrew Aramaic Abbreviation Disambiguation System,” *Journal of the American Society for Information Science and Technology*, 61:9 (2010), 1923–1932.
- Hockey, Electronic Texts = Susan Hockey, *Electronic Texts in the Humanities*, Oxford: Oxford University Press, 2000
- Ide-Véronis, TEI = Nancy Ide and Jean Véronis (ed.), *Text Encoding Initiative: Background and Context*, The Netherlands: Kluwer Academic Publishers, 1995
- Jockers, Macroanalysis = Matthew L. Jockers, *Macroanalysis: Digital methods and literary history*, University of Illinois Press, 2013.
- Karpathy, RNN = Andrej Karpathy, “The Unreasonable Effectiveness of Recurrent Neural Networks”, 2015 [<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>]
- Kestemont, Function Words = Mike Kestemont, “Function Words in Authorship Attribution. From Black Magic to Theory?”, *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, 2014, 59-66.
- Kirschenbaum, DH = Kirschenbaum, Matt, “What Is Digital humanities and What’s It Doing in English Departments?” (2011), [http://mkirschenbaum.files.wordpress.com/2011/01/kirschenbaum\\_ade150.pdf](http://mkirschenbaum.files.wordpress.com/2011/01/kirschenbaum_ade150.pdf)
- Koppel, Responsa = Moshe Koppel, “The Responsa Project: Some Promising Future Directions” in N. Dershowitz and E. Nissan (eds.), *Language, Culture, Computation: Essays Dedicated to Yaacov Choueka*, Springer-Verlag, Berlin, 2011
- Koppel-Schler, Stylistic Idiosyncrasies = M. Koppel and J. Schler, “Exploiting Stylistic Idiosyncrasies for Authorship Attribution,” *Proceedings of IJCAI’03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico, 2003

- Koppel-Schler, Authorship Verification = M. Koppel and J. Schler, "Authorship Verification as a One-Class Classification Problem," Proceedings of 21st International Conference on Machine Learning, July 2004, Banff, Canada, 489-495.
- Koppel et al., Unmasking = Moshe Koppel, Jonathan Schler and Elisheva Bonchek-Dokow, "Measuring Differentiability: Unmasking Pseudonymous Authors," Journal of Machine Learning Research 8 (2007), 1261-76.
- Liebeskind et al., Construction = Chaya Liebeskind, Ido Dagan and Jonathan Schler, "Semiautomatic Construction of Cross-Period Thesaurus," ACM Journal on Computing and Cultural Heritage 9, 4 (2016)
- Mahpod and Keller, Rashi Scripts = Shahar Mahpod and Yosi Keller, "Auto-ML Deep Learning for Rashi Scripts OCR", arXiv:1811.01290, 2018
- Mikolov et al., Efficient = Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," Neural Information Processing Systems 2013
- Mirkin, Agnon = Reuven Mirkin, עגנון במחשב [Hebrew], in: Kovetz Agnon 1 (1994), 319-344
- Moretti, Distant = Franco Moretti, Distant Reading, Verso, 2013.
- Moretti, Graphs = Franco Moretti, Graphs, Maps, Trees: Abstract Models for a Literary History, Verso, 2005.
- Munz-Manor, Computational Study = "Analog Piyyut in a Digital World: Towards Computational Study of Payytanic Literature", Jerusalem Studies in Hebrew Literature 32 (2021; Jubilee volume for Shulamit Elizur), 69-98.
- Netzer et al., Evaluating = Yael Netzer, David Gabay and Oren Hazai, "'This song is quite banal' – Evaluating Hebrew Lyrics," 11th Bar-Ilan Symposium on the Foundations of AI, 2011.
- Och-Ney, Systematic = Franz Josef Och and Hermann Ney, "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics 29, 1, 19-51, 2003.
- Pierazzo, Digital Editing = Elena Pierazzo, Digital Scholarly Editing, Routledge 2020.
- Raziel-Kretzmer and Shmidman, Mapping = Vered Raziel-Kretzmer and Avi Shmidman, "Computer-Aided Mapping of Palestinian Liturgical Rites as Represented by the Cairo Genizah Manuscripts," manuSciences '15 Conference, Chiemsee, Germany, September 2015. [Poster; available here: <http://tinyurl.com/pt9zr2r>]
- Rubin, Abbreviations = Aaron D. Rubin, "Abbreviations," Encyclopedia of Hebrew Language and Linguistics, Volume 1, Leiden and Boston: Brill, 2013, 1-4
- Rubinstein and Shmidman, Chronolect = Aynat Rubinstein and Avi Shmidman, "NLP in the DH pipeline: Transfer-learning to a Chronolect", Workshop on Natural Language Processing for Digital Humanities (NLP4DH 2021), Proceedings of the Workshop, 106-110
- Schoch, Data = Christof Schoch, "Big? Smart? Clean? Messy? Data in the Humanities," Journal of Digital Humanities, 2, 3 (2013)
- Schreibman et al., Companion = Susan Schreibman, Ray Siemens and John Unsworth (eds.), A Companion to Digital Humanities, Oxford: Blackwell, 2004
- Seker et al., AlephBERT = Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, Reut Tsarfaty, "AlephBERT: A Hebrew Large Pre-Trained Language Model to Start-off your Hebrew NLP Application With", arXiv:2104.04052
- Schorr et al., ViS-Á-ViS = Moshe Schorr, Oren Mishali, Benny Kimelfeld, Ophir Münz-Manor, "ViS-Á-ViS : Detecting Similar Patterns in Annotated Literary Text", IEEE Visualization Conference 2020
- Shmidman et al., Challenge Sets = Avi Shmidman, Josh Guedalia, Shaltiel Shmidman, Moshe Koppel, Reut Tsarfaty, "A Novel Challenge Set for Hebrew Morphological Disambiguation and Diacritics Restoration", Findings of the Association for Computational Linguistics: EMNLP 2020, 3316-3326

- Shmidman, Biblical Citations = Avi Shmidman, "Automatic Identification of Biblical Citations and Allusions in Hebrew Texts", in: #DHJewish - Jewish Studies in the Digital Age, ed. Michelle Chesner, et al., Berlin: DeGruyter, 2022 (in press)
- Shmidman et al., Identification = Avi Shmidman, Moshe Koppel, and Ely Porat, "Identification of Parallel Passages across a Large Hebrew/Aramaic Corpus", *Journal of Data Mining and Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, March 2018
- Shmidman et al., Nakdan = Avi Shmidman, Shaltiel Shmidman, Moshe Koppel, Yoav Goldberg, "Nakdan: Professional Hebrew Diacritizer," in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, July 2020, 197-203.
- Siemens-Schreibman, Literary = Ray Siemens and Susan Schreibman (eds.), *A Companion to Digital Literary Studies*, Oxford: Blackwell, 2013
- Suissa, Error Correction = Omri Suissa, "Optimizing OCR error correction of historical newspapers in Hebrew using neural networks and crowdsourcing", Ph.D. Thesis, Bar-Ilan University [Hebrew], 2019.
- Sutherland, Electronic = Katherine Sutherland (ed.), *Electronic Text: Investigations in Method and Theory*, Oxford: Clarendon Press, 1997
- Tsarfaty, Integrated = Tsarfaty, Reut, "Integrated morphological and syntactic disambiguation for modern hebrew," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, ser. COLING ACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, 49–54.
- Tsarfaty et al., Hebrew NLP = Reut Tsarfaty, Shoval Sadde, Stav Klein, Amit Seker, "What's Wrong with Hebrew NLP? And How to Make it Right", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations 2019*, 259-264.
- Tsvetkov and Wintner, MWE = Yulia Tsvetkov, Shuly Wintner, "Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources", *Computational Linguistics* 40, 2014/6, 449-468
- Widdows, Geometry and Meaning = Dominic Widdows, *Geometry and Meaning*, Stanford 2004
- Wintner, Morphological = Shuly Wintner, "Morphological Processing of Semitic Languages," in: Imed Zitouni (ed.), *Natural Language Processing of Semitic Languages*, Berlin and Heidelberg, 2014.