דוח סיום מענק בנושא: כינון מחקר ותשתיות דיגיטליים בשיתוף עם הספרייה הלאומית

מענק רב-שנתי 9432

ד"ר ענת בן-דוד, פרופ' אורן סופר (ז"ל), ד"ר ענת לרנר, פרופ' רעות צרפתי, פרופ' יורם קלמן וד"ר ורד זילבר-ורוד
המעבדה הפתוחה למדיה ומידע, האוניברסיטה הפתוחה | 31 בדצמבר 2020

המעבדה הפתוחה למדיה ומידע מרכינה ראש ומודיעה בצער ובאבל על פטירתו של עמיתנו פרופ' אורן סופר.

אורן היה ממייסדי המעבדה, חבר הנהלת המעבדה, וחוקר מוביל במעבדה. תחומי המחקר של אורן היו תיאוריית מדיה, היסטוריה של המדיה, ותקשורת דיגיטלית. המחקר שלו במעבדה שבו ניתח באמצעות כלים דיגיטליים עיתונות יהודית היסטורית מהמאה ה-19 גישר וקישר תקופות היסטוריות, גישות מתודולוגיות ודיסציפלינות מחקריות. אורן היה איש חזון, וחזונה של מעבדת המדיה והמידע מממש את תפיסתו על חשיבות החשיפה של סטודנטים במדעי החברה לכלי המחקר הדיגיטליים המעודכנים ביותר.

אורן יחסר לנו, חברי המעבדה, כחוקר, כמנהל ובעיקר כקולגה.

אנו נמשיך את עבודתו של אורן במחקר רב- תחומי ובין-תחומי ובחיזוק הקשר בין המחקר להוראה באוניברסיטה הפתוחה.

יהי זכרו ברוך

תוכן עניינים

תקציר מנהלים 3

2. דיווח על פעילויות 7

3. אתגרים 10

4. הישגים 11

5. מעקב והערכה 12

6. יכולות מקצועיות 12

7. משוב 13

נספח א: איורים 14

נספח ב: מכתב 19

תקציר מנהלים

באוקטובר 2016 הוגשה ליד הנדיב הצעת מחקר לפרויקט תחת הכותרת Establishing a Research and Infrastructure Collaboration between the Open University of Israel and the National Library of Israel.

המגישה, ד"ר ענת בן-דוד, וארבעה שותפים חברי סגל בכיר באוניברסיטה הפתוחה, ביקשו לקדם פיתוח של מחקר דיגיטלי וכלים טכנולוגיים חינוכיים בתחום מדעי החברה החישוביים, תוך שימוש בנתונים ובאוספים דיגיטליים של הספרייה הלאומית.

חזון הפרויקט היה לא רק להרחיב את המידע מהאוספים על ידי הנגשת המידע, אלא גם להגדיל את ערכם לקהילת המחקר, כמו גם לציבור הרחב.

מטרת הפרויקט המחקרי היתה לבנות פרוטוטיפים לממשקי מחקר שיאפשרו חיפוש מתקדם על פני סוגים שונים של אוספי מדיה דיגיטליים (ארכיון רשת, אודיו/וידאו, עיתונות מקוונת והיסטורית), באופן שמצד אחד יאפשר לשלוף את המידע הרלוונטי בכל אוסף עבור שאלת המחקר, ומצד שני ייקח בחשבון את האלמנטים הייחודיים של כל מדיום לשאלת המחקר.

הפרויקט כלל חמישה מסלולים (ראו איור 1) – שלושה שהתמקדו באוספים דיגיטליים שונים, (ארכיון הרשת, ארכיון עיתונות היסטורית ועכשווית, ואוספי וידאו ואודיו), ושניים שעסקו בפיתוח אלגוריתמים וטכנולוגיות למידה עבור ניתוח נתוני עתק (פיתוח אלגוריתם זיהוי ישויות (Named Entity Recognition) הראשון בעברית, ופיתוח מעבדת דאטה לסטודנטים).

במהלך הפרויקט התמודד צוות המחקר בכל אחד מרכיבי הפרויקט עם אתגרים חישוביים, אינפורמטיביים או טכנולוגיים שונים. לסיכום הפרויקט, אנו גאים לומר שלא רק שפיתחנו מתודולוגיות, ממשקים ושיטות עבודה מומלצות לכל סוג של אוסף דיגיטלי, אלא שגם תרמנו למחקר התיאורטי בתחום מדעי החברה הדיגיטליים בכך שהצפנו שאלות רלוונטיות למחקרים היסטוריים וחברתיים העושים שימוש באוספים מסוגים אלו.

תוצרי הפרויקט משמעותיים ויכולים לשמש גם כפרוטוטיפים שייבחנו לצורך פיתוח נוסף על ידי הספרייה הלאומית (NLI), למחקרים דיגיטליים גנריים ופתוחים ולטכנולוגיות חינוכיות לטובת קהילת המחקר ותלמידי האוניברסיטה הפתוחה.

איור 1: חמשת רכיבי הפרויקט

להלן דוח הפעילות לסיום הפרויקט שנפרס על פני כמעט ארבע שנים (ינואר 2017 עד אוקטובר 2020), בדגש על שנת הפעילות הרביעית 31/10/2019–31/10/2020.

דוח הפעילות מבוסס על דוח מילולי זה שמוגש יחד עם דוח כספי.

<u>הדוח המילולי</u> מפרט את שבעת ההיבטים הבאים, עליהם התבקשנו לדווח:

1. **עדכון לגבי התקדמות הפרויקט**: תארו את המצב הנוכחי עם סיום תמיכת יד הנדיב בפרויקט. במה היה יישום הפרויקט שונה מהתכנון הראשוני ומה היו הגורמים לשינוי?
2. **דיווח על פעילויות**: מהן 3 הפעילויות המרכזיות שבוצעו במסגרת הפרויקט? האם במסגרת פעילויות אלו שותפו תוצרים כלשהם עם הציבור? (פרסומים, ידע, תכניות לימוד, כנסים וכו')
3. **אתגרים**: מהם השינויים המרכזיים שחלו בביצוע הפרויקט? ממה נבעו? כיצד השפיעו השינויים על המשך הפרויקט?
4. **הישגים**: מהם ההישגים המשמעותיים ביותר של הפרויקט?
5. מעקב והערכה: מהם המנגנונים שייצרתם כדי לעקוב ולהעריך את הפרויקט? שתפו אותנו בנתונים משמעותיים
6. **יכולות מקצועיות**: אלו יכולות מקצועיות נבנו בארגון במסגרת הפרויקט? האם אתם מוצאים ליכולות אלה ערך מעשי בפעילויות אחרות בארגונכם? האם היתה לתמיכת הקרן השפעה מעבר לכספי המענק?
7. **משוב**: נשמח לקבל משוב לגבי תהליך העבודה והיחסים עם יד הנדיב וכן לקבל כל הצעה לשינוי/ לשיפור

עבור כל היבט ניתנת סקירה כללית קצרה ולאחר מכן פירוט היבט זה בכל אחד מחמשת המסלולים.

<u>הדוח הכספי</u> של השנה הרביעית מציג הוצאות בסכום של 732,933 ₪.

<u>הדוח הכספי המסכם</u> מציג הוצאות על סך 2,894,000 ₪ שהן כ-100% מהתקציב הכולל המתוכנן.

כידוע, תכנית העבודה לשנה הרביעית לוותה גם בהצעת תקציב חדשה ובאוגוסט 2019 קיבלנו אישור להארכת הפרויקט בשנה נוספת (שנה רביעית), כדי לסיים את הפרויקטים: זיהוי ישויות (פרופ' רעות צרפתי) ומעבדת סטודנטים (פרופ' יורם קלמן). כמו כן תוכנן בשנה הרביעית המשך פיתוח פרויקט העיתונות ההיסטורית לאחר הגעה להבשלה בתהליך עיבוד הטקסטים. במסלולי ארכוב האינטרנט וניתוחי אודיו/וידאו המשכנו בעבודה מצומצמת בהתאם לקצב ניצול התקציב המפורט במסלולים המקבילים.

גם בשנה הרביעית היו הישגי הצוותים הם רבים ומשמעותיים. השנה הרביעית עמדה בסימן ניתוח ויישום מחקרי בשני המסלולים העיקריים:

1. מעבדת דאטה לסטודנטים
2. זיהוי ישויות בעברית

בכל מסלולי המחקר עשינו בדיקות היתכנות לגבי המתודולוגיות החדשות ואלו זכו לפרסום בקהילות המחקר השונות, כפי שיפורט בהמשך.

בזמן כתיבת שורות אלו, הלך מאיתנו בטרם עת פרופ' אורן סופר עקב מחלה קשה. אורן לא הספיק לקרוא את הדוח והקטעים שמתייחסים למסלול המחקר של העיתונות ההיסטורית נכתבו על ידי צוות המעבדה.

בברכה,

ד"ר ענת בן-דוד, פרופ' אורן סופר (ז"ל), ד"ר ענת לרנר, פרופ' רעות צרפתי, פרופ' יורם קלמן וד"ר ורד זילבר-ורוד.

1.עדכון לגבי התקדמות הפרויקט

תארו את המצב הנוכחי עם סיום תמיכת יד הנדיב בפרויקט. במה היה יישום הפרויקט שונה מהתכנון הראשוני ומה היו הגורמים לשינוי?

1.1. מעבדת דאטה לסטודנטים

מעבדת הדאטה לסטודנטים עסקה בנושא המיומנויות הנדרשות לניתוח דאטה (Data Analytics).

בשנה האחרונה פותח כלי הניסוי על ידי חברת BigBI תוך קשר ישיר עם אורלי וייסר, דוקטורנטית באוניברסיטת בן-גוריון בנגב בהנחייה משותפת של פרופ' יורם קלמן. אורלי היתה אחראית יחד עם פרופ' יורם קלמן על המחקר. במהלך השנה נמסרו גרסאות על ידי החברה, הן עברו בדיקות איכות שהוחזרו להמשך שיפור הכלי וחוזר חלילה. בשלב הסופי נערכו פיילוטים עם צוות המעבדה לקראת העמדת הכלי לטובת הניסוי. במקביל, הוכנו כלי המחקר וחומרי ההדרכה למשתתפי המחקר.
הניסוי עצמו, בהשתתפות סטודנטים, החל בנובמבר 2020 והוא נמצא בעיצומו (ראו צילומי מסך המתעדים את הניסוי באיור 1 בנספח א). בסופו של הניסוי יתבצע ניתוח מעמיק הן של השאלונים שמילאו הסטודנטים והן באמצעות קריאת הלוגים שתועדו בעת עבודת הסטודנטים על הכלי.

1.2. זיהוי ישויות בעברית

מסלול זה עסק במתייג ומסווג ישויות לעברית. מספר תסריטים פותחו במהלך הפרויקט:
א. pipeline עם סגמנטציה
ב. Joint עם סגמנטציה
ג. Hybrid architecture

פרטים נוספים ניתן לקרוא בפוסטר שפורסם השנה (ראו איור 2 בנספח א).

במהלך הפרויקט שקלנו לנהל תיוג המונים של קורפוס חופשי כגון ויקיפדיה, אולם זה לא יצא לפועל. חלק עקב התפרצות ה-COVID-19 וחלק בגלל אתגרים טכניים של רוחב פס.

1.3. עיתונות היסטורית

מסלול העיתונות ההיסטורית עסק בשיפור מנוע OCR של עיתונות יהודית מהמאה ה-19 ובמחקר חישובי באמצעות אלגוריתמים שונים לניתוח שיח. הפרויקט הצליח מאוד ברמה הטכנית-תשתיתית (ראו נספח ב, מכתב מפרופ' ירון צור) וזאת איפשרה להשיג תוצאות בניתוח החישובי. להלן שלבי הפרויקט מתחילתו ועד סופו.

1. פיילוט הצפירה: בתחילת הפרויקט ביצענו פיילוט שהוקדש לטיוב ה-OCR של עיתון הצפירה שלווה בניסוי של החלת שיטות מחקר חישוביות על הטקסט המתקבל (קישור למצגת הסיכום של שלב זה). תוצאות הניסוי אף התפרסמו בכתבי עת מדעיים.
2. Scaling up: בשלב השני, בחנו את מודל ה-OCR שאומן על עיתון הצפירה על מגוון של עיתונים, בעיקר מן המאה ה-19, באמצעות סקר מקיף. בגלל היקף הפרויקט, שלב זה כלל גם שיתוף פעולה עם פרופ' ירון צור ומימון נוסף לפרויקט. בתוך כך אימנו מודל כללי משופר שמתאים לעיתונות העברית של המאה ה-19. בסופו נבנה קורפוס השוואתי שכולל עשור מעיתוני 'המגיד' ו'המליץ', בצד העשור המקביל מן 'הצפירה'.
3. פייפליין: בשלב השלישי והאחרון השלמנו אוטומציה של תהליך הטיוב וארזנו 'פייפליין' - חבילת תוכנה שבאמצעותה יכול כל מתכנת, בכמה פקודות פשוטות, להריץ את התהליך. הפייפליין מקבל את קובצי העיתונות ההיסטורית כפי שנמסרו מן הספריה (עם ניתוח של Olive Software), משמר את המידע המבני והלוגי יקר-הערך אך מזהה מחדש את השורות ואת הטקסט שבהן באיכות משופרת. הטקסט מיוצא לפורמטים טקסטואליים שונים שמאפשרים שימושים שונים ומחקר בשיטות דיגיטליות של ניתוח שפה וטקסט. הפייפליין עודכן במהלך אוגוסט 2020 כדי להכיל שיפורים חדשים מבית טרנסקריבוס, ובעיקר הוספה של שימוש במודל שפה, ששיפרה עוד יותר את אחוזי דיוק ה-OCR.
4. כמו כן נעשה ניסיון ראשוני, שלא צלח עדיין, להוסיף פורמט הניתן לתצוגה על ידי מערכת ורידיאן שמופעלת עבור ה-JPRESS בספרייה הלאומית.

1.4. ארכיון הרשת

מסלול ארכוב האינטרנט רשם הישגים רבים בתחום קידום מתודולוגיות דיגיטליות לחקר ארכיוני רשת. פיתחנו כלי מחקר ומסגרות תיאורטיות ומתודולוגיות לחקר ארכיוני רשת, תוך שימת לב מיוחדת למגבלות הטכנולוגיות והמשפטיות בגישה לארכיונים מסוג זה, הן בארכיונים מוסדיים המנוהלים על ידי ספריות לאומיות, והן בגישה למידע היסטורי מרשתות חברתיות ומידע מקוון אחר. קשיי גישה אלה, הובילו אותנו בסופו של דבר לעבוד עם מסדי נתונים אחרים ממסד הנתונים שהונגש לנו על ידי הספריה הלאומית. עם זאת, כיוון שבעיית הגישה אינה ייחודית לישראל,

הכלים והמתודולוגיות האלטרנטיביות שפיתחנו – Retrospective web archiving ו-Counter-Archiving, פתחו אפשרויות מחקר חדשות עבור היסטוריונים של הרשת, שלא התאפשרו עד כה.

1.5. אודיו/וידאו

בתכנון המקורי, חזון מסלול האודיו/וידאו היה הנגשת האינפורמציה המצויה באודיו ווידאו באמצעות תמלול אוטומטי של התוכן הנאמר. משימה זאת בוצעה על ידי פיתוח כלי תמלול שעומד לרשות הציבור הרחב: http://hebrew-transcriber.online. ואולם, יישום הפרויקט היה מקיף יותר מהתכנון המקורי. הערך המוסף של מחקר האודיו-וידאו היה בפיתוח מתודולוגיה לחילוץ מידע נוסף על זה שיש בתמלול התוכן ופיתחנו שיטות לניתוח פרמטרים אקוסטיים ומדידות כמותיות, בעיקר בחקר דיאלוגים.

דווקא במהלך השנה האחרונה הצלחנו לגשר על פער מסוים בין המתודולוגיה שפותחה בשנים הקודמות, לבין החזון של העשרת אוצרות התרבות הקוליים וחילוץ המידע הכלוא בהם. הגישור נעשה באמצעות חבירה לקבוצות מחקר בתחום מדעי הרוח והחברה הדיגיטליים ((Digital Humanities and Social Sciences (DHSS).

בשנה הרביעית לפרויקט, בעוד אנחנו בוחנים תרחישי מחקר חדשניים וממשקים לאוספים דיגיטליים רב-מודאליים, כולל שמע ווידאו, שנאצרו על ידי הספרייה הלאומית בישראל, נהיו אוספי השמע והווידאו רלוונטיים יותר מתמיד. במהלך 8-10 החודשים האחרונים, בעקבות התפרצות מגפת הקורונה, הפכו שיחות וידאו (בעיקר באמצעות פלטפורמת Zoom) לכלי תקשורת מרכזי, הזירה החינוכית מתרכזת כעת בכלי למידה מבוססי דיבור, ולא יעבור זמן רב עד שהיא תחפש כלים אנליטיים לאובייקטי למידה אלה. גם הדיסציפלינה האקדמית של oral history, שמתבססת על ראיונות בעל-פה מחייבת פרקטיקות חדשות, של ראיונות מרחוק בעקבות מגבלות המגפה.

לסיכום, יישום הפרויקט היה שונה מהתכנון המקורי ומבחינות רבות הרחיב את החזון המקורי. ובנוסף, אירועים גלובליים מיצבו את אופנות האודיו והווידאו בחזית והישגי הפרויקט יתרמו לא רק לאוצרות רוח בספרייה הלאומית, אלא לעיבוד אוספי שמע בכלל.

2.דיווח על פעילויות

מהן 3 הפעילויות המרכזיות שבוצעו במסגרת הפרויקט? האם במסגרת פעילויות אלו שותפו תוצרים כלשהם עם הציבור? (פרסומים, ידע, תכניות לימוד, כנסים וכו')

להלן התייחסות לכל אחד מרכיבי הפרויקט בנוגע לסעיף הפעילויות:

2.1 מעבדת דאטה לסטודנטים

שלוש הפעילויות המרכזיות הן:

א. ביצוע מחקר דלפי בינלאומי העוסק במיומנויות הנדרשות לניתוח דאטה (Data Analytics). המחקר נערך בשלושה סבבים, בהשתתפות עשרות מומחים מרחבי העולם. התוצאות מראות כי קיים קונצנזוס בין המומחים לגבי מיומנויות שהן חיוניות לביצוע ניתוח מידע ואת רשימת אותן מיומנויות.

ב. לצורך יישום הממצאים של מחקר דלפי, נבחן הקשר בין המיומנויות שאופיינו כחיוניות לניתוח מידע לבין ביצוע בפועל של משימות ניתוח מידע, באמצעות אבטיפוס של מעבדת הסטודנטים. בהתאם לתוצאות מחקר הדלפי, נבחרה רשימה מצומצמת של מיומנויות לבדיקה ופותח כלי מחקר באמצעות התוכנה *Sparktify* על ידי חברת *BigBI*. הניסוי כולל דיווח של הנבדקים על התקדמותם ועל התוצאות אליהן הם מגיעים, ועוד.

ג. ועדת ההיגוי להקמת מעבדת סטודנטים שהוקמה במהלך הפרויקט החליטה להקים צוות מקצועי מצומצם שיבחן את הצרכים וימליץ על מאפייני הפתרון, בדגש על המפרט הטכני הנדרש. הצוות המצומצם כלל אנשי מקצוע ממספר גופים באוניברסיטה הפתוחה (האו"פ), בעלי ניסיון טכני ופדגוגי, המכירים את מערכות המיחשוב של האו"פ. הצוות בחן את מאפייני הכלי הנדרש לביצוע ניסוי המעבדה. אנו רואים בכך שלב ראשון לקראת הקמת מעבדת הסטודנטים ושילובה במערך ההוראה של האו"פ.

2.2 זיהוי ישויות בעברית

הפעילויות העיקריות כללו:

א. יצירת קורפוס מתויג של עיתון 'הארץ', ששימש בעבר כקורפוס תשתיתי לאימון של כלי NLP אחרים בעברית.

ב. הרצת סט של מודלים מאומנים על הקורפוס. בניתוח התוצאות נמצאו יתרונות מובהקים לשימוש בפירוק המורפולוגי עבור מידול NER.

ג. קוד לזיהוי ישויות בעברית. רק לאחר שהמאמר שנמצא בתהליך שיפוט יתקבל, נוכל לפרסם את הקוד בפלטפורמה שיתופית.

פרסומים בכתבי עת:

Bareket, D., & Tsarfaty, R. (2020). Neural Modeling for Named Entities and Morphology (NEMO^2). *arXiv preprint arXiv:2007.15620*. https://arxiv.org/abs/2007.15620

מאמר נוסף נמצא בתהליך ביקורת עמיתים לכתב העת TACL https://transacl.org/index.php/tacl

פרסומים בדברי כנס (Proceedings):

המחקר הוצג בכנס ISCOL וכן בהרצאות בסמינר באוניברסיטת בר אילן (ראו איור 2 בנספח א).

Bareket, D., & Tsarfaty, R. (2020). Neural Modeling for Named Entities and Morphology (NEMO^2). *The Annual Meeting of the Israeli Seminar on Computational Linguistics (ISCOL 2020)*, September 8, 2020. https://iscol2020.github.io/

2.3 עיתונות היסטורית

הפעילויות העיקריות כללו:

א. אין ספק שהצלחת מסלול העיתונות ההיסטורית היתה הודות לפיתוח הפייפליין (ראו לעיל סעיף 1.3 ומכתב מפרופ' ירון צור ומר איל מילר בנספח ב). הפייפליין הינו קוד שפותח במעבדה והוא זמין להורדה ושימוש, כולל הוראות הרצה, במערכת גיטהאב:
https://github.com/omilab/historical_press/tree/master/OCR_Pipeline
הפייפליין בנוי משלשה סקריפטים עוקבים אך נפרדים:

- Legacy_to_tkbs_format_converter.py
ממיר את המידע המבני (החלוקה לאזורים וסידורם בדף) של קובצי PRXML שאותם קיבלנו מהספרייה הלאומית לפורמט PAGEXML הנדרש להזנה במערכת טרנסקריבוס.
- Tkbs_uploader.py
מעלה קבצים לשרת של טרנסקריבוס, מבצע קריאת שורות וטקסט ומייצא את התוצרים בפורמט PAGEXML
- tkbs_exporter.py
ממיר את תוצרי הקריאה המטויבת של טרנסקריבוס לפורמטים נוחים לניתוח ומחקר: פורמט טקסט פשוט, ופורמטים מובנים של קבצי XML-TEI ו-TSV (פורמט טבלאי).

ב. מענקים:
  - מענק ISF ציוד מדעי עבור חבר סגל "אמצע הדרך" לפרופ' אורן סופר, ע"ס 141,000 ₪.
  - ציון טוב מאוד של מענק ISF אישי שכתבו פרופ' אורן סופר וד"ר זף סגל.
  - מענק ע"ס 100,000 ₪ מטעם רשות המחקר באוניברסיטה הפתוחה בנושא Digital Access to Textual Cultural Heritage in Hebrew – חוקרים ראשיים: פרופ' אורן סופר וד"ר זף סגל.

ג. כנס וסדנה בינ"ל בנושא תוכננו במהלך השנה השלישית והיו אמור להתקיים במשך יומיים ביולי 2020. תכנית הכנס היתה סגורה עם מרצים אורחים מחו"ל, אולם עקב התפרצות מגפת הקורונה נאלצנו לבטל את הכנס.

פרסומים בכתבי עת:

Segal, Z. M. & Soffer, O. (2020). One journal, one decade, 3,797,592 words. *Journal of Jewish Studies.*

Segal, Z. M. & Soffer, O. (2020). From Weekly to Daily: Computational Analysis of Periodical Time Cycles. *Journalism Studies*, 1-21. https://doi.org/10.1080/1461670X.2020.1807394

פרסומים בדברי כנס (Proceedings):

Segal, Z. M. (2021). Constructing the Modern Jewish “Present”: Computational Analysis of Periodical Time Cycles in HaTzfira”, To be presented during the international conference “#DHJewish - Jewish Studies in the Digital Age”, C^{2}DH, University of Luxemburg, 11-13 January, 2021.

Segal, Z. M. (2020). “Constructing the Modern Jewish “Present”: Computational Analysis of Periodical Time Cycles in HaTzfira”, in: Software for the Past (SfP): Digital Technologies to Study the Past and Present, Kinneret, December 2020 (lecture).

2.4 ארכיון הרשת

הפעילויות הבולטות של פרויקט ארכוב האינטרנט היו:

א. כנס ארכוב אינטרנט בינלאומי, בשיתוף עם הספריה הלאומית, פתוח לקהל הרחב.

ב. פרסומים אקדמיים – ובעיקרם מאמרים מתודולוגיים המדגימים שיטות חדשות לחקר ארכוב האינטרנט כמקרי בוחן.

ג. פרסום כלי מחקר (למשל, *https://github.com/omilab/internet-archive-link-extractor*) וממשקי משתמש לדוגמה (ראו איור 3 בנספח א).

פרסומים בכתבי עת:

Ben-David, A. (2020). Counter-archiving Facebook. *European Journal of Communication*, https://doi.org/10.1177%2F0267323120922069

2.5 אודיו/וידאו

שלוש הפעילויות המרכזיות שבוצעו במסגרת המסלול:

א. שותפות בינלאומית נוצרה עם שלושה חוקרים מארה"ב, מומחים בתחום ניתוח דיאלוגים ויחסים בין דוברים בדיאלוג: Prof. Julia Hirschberg (Columbia University), Prof. Rivka Levitan (Brooklyn College), and Andreas Weise (City University of New York (CUNY)). במהלך השנה פורסם מאמר משותף בכתב העת *Journal of Phonetics*.

ב. ערב עיון שמסכם את פעילות מסלול האודיו-וידאו התקיים ב-19 באוקטובר 2020. ערב העיון נשא את הכותרת: *Interdisciplinary day on audio and video collections*
קישור לתכנית האירוע: https://www.publicators.com/app/dms.asp?ms_id=26437

- ד"ר זילבר-ורוד הוזמנה להיות מרצה אורחת בסדנה בינלאומית שהתקיימה בחסות פרויקט *CLARIN* של האיחוד האירופי (*CLARIN - European Research Infrastructure for Language Resources and Technology*) ראו www.clarin.eu. נושא ההרצאה היה:

Enriching audio databases with information hidden in the acoustic signal, "Speech, Voice, Text, and Meaning" workshop about Oral History and Technology, 29.10.2020.

פרסומים בכתבי עת:

Weise,A., Silber-Varod, V., Lerner,A., Hirschberg, J., Levitan, R. (2020). Entrainment in spoken Hebrew dialogues. In: J. Pardo, E. Pellegrino, V. Dellwo, and B. Möbius (Eds.), Special Issue on *Vocal Accomodation in Speech Communication*, *Journal of Phonetics, Vol. 83*. https://doi.org/10.1016/j.wocn.2020.101005

Lerner, A., Silber-Varod, V., Carmi, N., Guttel, Y., & Allouche Omri (Forthcoming). Modeling the dynamics of acoustic gaps between speakers during Business-to-Business sales calls. International Journal of Big Data Intelligence (IJBDI).

Silber-Varod, V., Malayev, S., & Lerner, A. (2020). Positioning Oneself in Different Roles: Structural and Lexical Measures of Power Relations between Speakers in Map Task Corpus. *speech communication, 117*: 1-12. https://doi.org/10.1016/j.specom.2020.01.002

פרסומים בדברי כנס (Proceedings):

Silber-Varod, V., Amit, D., Lerner, A. (2020). Tracing changes over the course of the conversation: A case study on filled pauses rates. *Proc. 10th International Conference on Speech Prosody 2020* (pp. 754-758), DOI: 10.21437/SpeechProsody.2020-154.

Silber-Varod, V., Lerner, A., Carmi, N., Amit, D., Guttel, Y., Orlob, C., & Allouche, O. (2019). Computational modelling of speech data integration to assess interactions in B2B sales calls. IEEE 5th International Conference on Big Data Intelligence and Computing (IEEE DataCom 2019) , *pp. 152-157. DOI 10.1109/DataCom.2019.00031*

3.אתגרים

מהם השינויים שחלו בביצוע הפרויקט? ממה נבעו? כיצד השפיעו השינויים על המשך הפרויקט?

להלן התייחסות לכל אחד מרכיבי הפרויקט בנוגע לסעיף ניהול השינויים:

3.1. מעבדת דאטה לסטודנטים

השינוי העיקרי היה העמקת הבסיס המחקרי עקב שילובה של אורלי וייסר לפרויקט, דבר שאפשר חקירה מעמיקה יותר הן של המיומנויות הנדרשות, הן של הקשר בין המיומנויות האלו לבין היכולות של סטודנטים לבצע מטלות של אנליטיקות. בזכות זה המחקר בתחום ממשיך גם מעבר לתאריך הסיום של מימון הפרויקט, ובאוניברסיטה התבצעה היערכות כדי לבנות בעתיד מערכת של האוניברסיטה שתיישם את ממשאי המחקר.

בנוסף, בשנה האחרונה התמודדנו עם התפרצות מגפת הקורונה. במסלול מעבדת הסטודנטים היה הצורך לשנות את מתווה הניסוי שתוכנן להתבצע במעבדה למגבלות *Covid19*. הניסוי נבנה מחדש באופן שיאפשר לבצע אותו מרחוק. ביצוע הניסוי בפועל נערך במהלך נובמבר 2020 וניתוח התוצאות יתבצע לאחר מכן.

3.2. זיהוי ישויות בעברית

זהו פרויקט מורכב ורחב היקף, הן ברמת תשתיות השפה והן ברמה החישובית. במהלך הפרויקט ערכנו ניסויים רבים שהניבו תוצאות רבות שאותן היה צריך לנתח. הגענו לתוצאות מהימנות שדורשת תשומת לב והבנה בפרטים טכניים ותאורטיים רבים.

3.3. עיתונות היסטורית

היו שינויים בדרישות התקציב עקב התבססות על שירות של תוכנות חיצוניות בתשלום: הפייפליין (ראו סעיף 2.3) הוא קוד חופשי, אך החל מספטמבר 2020 השימוש בתוכנת טרנסקריבוס (שלב 2 של הקוד) הוא שירות בתשלום, הזמין בהוזלות מיוחדות למנויים ולמוסדות חברים בקואופרטיב READ. הודות למענק ISF לציוד מדעי של חוקר באמצע הדרך שקיבל פרופ' אורן סופר השגנו את התקציב הנדרש לרכישת מנוי לשימוש בטרנסקריבוס שישמש אותנו בשנה הקרובה (2021).

בהמשך לכך, חלק 2 של קוד הפייפליין מותאם להרצה של מספר גליונות מוגבל, והרצה תעשייתית של מעבר למאות גליונות בודדים אורכת זמן רב. לכן, כשתתקבל החלטה לטייב את ה-OCR של JPRESS מומלץ להחליף את שלב 2 בתקשורת ישירה עם צוות טרנסקריבוס, שעם קבלת האוסף בשלמותו יריץ עליו את הפעולות על שרת מקומי ובמהירות.

3.4. ארכיון הרשת

האתגר המשמעותי של פרויקט ארכוב האינטרנט היה הקושי לבצע עיבוד חישובי של נתוני ארכיון האינטרנט הישראלי שהונגש לנו לטובת המחקר, בשל אילוצים תשתיתיים שנבעו מהצורך לשמור את הנתונים בשרתים של הספריה. סביבת העבודה שהונגשה עבור הפרויקט לא התאימה לעבודה החישובית המורכבת שתכננו לבצע על נתונים אלה. לכן, מרבית הפעילות המחקרית התמקדה בניתוח נתונים שהיו זמינים ברשת הפתוחה, הן מארכיון האינטרנט והן מאתרי אינטרנט ורשתות חברתיות.

3.5. אודיו/וידאו

האתגר הגדול לאורך השנים היה הפעלת כלי התמלול האוטומטי שפיתחנו במעבדה במסגרת הפרויקט.

בשנה הראשונה השתמשנו במנוע הזיהוי של חברת גוגל (Google Cloud speech to text API). הכלי אינו חינמי, אך איפשר כמות מסוימת של הרצות בחינם. במהלך השנה הראשונה קיבלנו פידבקים חיוביים ממשתמשים רבים אולם הכלי הושבת מפעם לפעם והיה צורך בתחזוקה שוטפת שלו. לאחר שנה, עברנו להשתמש במנוע זיהוי דיבור בעברית של חברת קונמג'י https://www.conmagi.com שניתן לנו בחינם. לאחר שנה, ובעקבות פרוץ הקורונה, הודיע לנו מייסד החברה שהם לא יוכלו להמשיך במתכונת החינמית. בעקבות כך, קיבלנו תקציב ממחלקות שונות באו"פ להתחבר מחדש למנוע של גוגל באמצעות שירותי מחב"א (המרכז החישובי הבין אוניברסיטאי).

המתמלל הוא כלי ייחודי בנוף האינטרנט הישראלי והשירות שהוא נותן לצורכי מחקר והתרשמות נועד לקדם את המודעות לחשיבות שיש בפיתוח טכנולוגיות שפה לעברית. לצערנו, אם לא יימצא תקציב נוסף, ניאלץ לסגור את הכלי בסוף 2020.

4.הישגים

מהם ההישגים המשמעותיים ביותר של הפרויקט?

להלן התייחסות לכל אחד מרכיבי הפרויקט בנוגע לסעיף ההישגים והקשיים:

4.1. מעבדת דאטה לסטודנטים

- ההישג המשמעותי ביותר הוא הבנה מעמיקה יותר של הקשר בין מיומנויות של סטודנטים ועובדים בתעשייה, לבין היכולת שלהם להתמודד עם אתגרים עכשויים בתחום של אנליטיקות מידע בכלל ואנליטיקות מידע שקשורות לטקסט באופן ספציפי.
- בנוסף, יש בידינו אבטיפוס של כלי שנוכל להעמיד לרשותם של חברי סגל שמעוניינים להשתמש במאגרי מידע עצומים כמו אלו שיש בספרייה הלאומית כדי להכין מטלות אנליטיקת מידע עבור סטודנטים בדיסציפלינות שונות במדעי המחשב, במדעי החברה ובמדעי הרוח.
- מחקר דלפי: בשנים הקודמות ערכנו מחקר דלפי העוסק במיומנויות הנדרשות לניתוח נתונים (Data Analytics).
- מעבדת הסטודנטים: ועדת ההיגוי להקמת מעבדת סטודנטים בחרה בכלי BigBI Studio (לשעבר Sparktify) של חברת BigBI (סטארטאפ ישראלי בהנהלת מר רביב קנולר) לביצוע ניסוי לבחינת מיומנויות ניתוח נתונים של סטודנטים. בהמשך ל–POC המוצלח שהתבצע בשנה הקודמת, בשנה זו תכננו ובנינו ניסוי מעבדה שיתבצע כשלב ראשון לפני הקמת מעבדת הסטודנטים ושילובה במערך ההוראה של האו"פ.
ניסוי המעבדה החל ויימשך במהלך שנת הלימודים תשפ"א 2021. על בסיס תוצאות מחקר הדלפי, בחרנו רשימה מצומצמת של מיומנויות לבדיקה בניסוי. המשתתפים בניסוי מתבקשים למלא שאלוני מיומנויות ושאלוני אישיות. בנוסף, הם מתבקשים לבצע משימת ניתוח נתונים שבנינו בעזרת הכלי BigBI Studio. הפעילות תירשם בקובצי לוג של המערכת. בסיום הניסוי, קובצי הלוג, שאלוני המיומנויות והאישיות ותוצאות ביצוע המשימה ינותחו ע"י צוות המעבדה.

בנוסף, עבור הדוקטורנטית אורלי וייסר השנה היתה רבת הישגים, הודות למחקר שהיא ניהלה במסגרת פרויקט יד הנדיב.

- במהלך השנה האחרונה, אושרה המועמדות שלה ללימודי דוקטורט באוניברסיטת בן-גוריון בנגב.
- היא זכתה בשתי מלגות תחרותיות ויוקרתיות.
  - האחת מלגת נשיא המדינה 2020 למצויינות וחדשנות מדעית בנושא: "מיומנויות ניתוח מידע ככלי להעצמת אזרחים - כלב השמירה של העיתונות, הדמוקרטיה וריבונות העם"
  - השנייה ממשרד המדע.
- הגשת מאמר לפרסום בז'ורנל Q1. המאמר מסכם את ניסוי הדלפי והמסקנות שעלו ממנו.

4.2. זיהוי ישויות בעברית

- סגירת גרסה לקורפוס החדש ל-NER בעברית, עליה אומנו המודלים והורצו כלל הניסויים.
- בניית מערך ניסויים לבחינה של יחסי מורפולוגיה ו-NER בעברית במודלי רשתות נוירונים.
- פיתוח שיטה לביצוע של NER עם התחשבות בגבולות מורפולוגיים: שיטה חדשה לביצוע היברידי של פירוק מורפולוגי וזיהוי ישויות.

4.3. עיתונות היסטורית

כאמור, ההישגים העיקריים הם:

- פיתוח הפייפליין כתשתית למחקר שיכול להתפתח בתחום מדעי החברה הדיגיטליים.
- מחקר ענף בכלים חישוביים ומתוך תפיסת קריאה רחוקה (distance reading), שאישושה ידע קיים וחשפה ידע חדש שלא נחקר עד כה.

4.4. ארכיון הרשת

ההישג המחקרי המשמעותי ביותר של הפרויקט הוא פיתוח מתודולוגיה לארכוב האינטרנט בדיעבד, באופן שמאפשר לשחזר אירועים לאומיים ובינלאומיים שלא אורכבו בזמן אמת. מתודולוגיה זו יכולה לשמש הן מוסדות שימור לאומיים, והן קהילות מחקר העוסקות בהיסטוריה של האינטרנט ושימורה.

4.5. אודיו/וידאו

ההישגים העיקריים הם:

- חילוץ ידע מגוון מקובצי אודיו ווידאו הנותן ערך מוסף על פני תמלול אוטומטי של הדיבור (ראו איור 4 בנספח).
- מתודולוגיה חדשה לניתוח דיאלוגים ולהשוואות בין דיאלוגים ובכך הנחנו את היסוד לעיבוד של בסיסי נתונים קוליים גדולים.

5. מעקב והערכה

מהם המנגנונים שייצרתם כדי לעקוב ו/או להעריך את הפרויקט? שתפו אותנו בנתונים משמעותיים

מעבר לניהול השוטף של הפרויקט על ידי מנהלת המעבדה, ד"ר ורד זילבר-ורוד, שעקבה אחרי פעילות עוזרי המחקר ובשיתוף פעולה עם החוקרים הראשיים, החלטנו באפריל 2019 לכתוב את רשימת התוצרים הצפויים בפרויקט. התוצרים הצפויים חולקו ל-4 קטגוריות: פרסום, כלים, קוד ודאטה והחוקרים הראשיים ציינו תוצרים בכל אחת מהקטגוריות. הדבר אפשר לנו לנהל מעקב אחרי מטרות כל מסלול מחקרי.

אנחנו שמחים לציין שכמעט כל התוצרים הושגו במלואם, כפי שבא לביטוי בדוח זה.

6. יכולות מקצועיות

אלו יכולות מקצועיות נבנו בארגון במסגרת הפרויקט? האם אתם מוצאים ליכולות אלה ערך מעשי בפעילויות אחרות בארגונכם? האם היתה לתמיכת הקרן השפעה מעבר לכספי המענק?

להלן התייחסות לכל אחד מרכיבי הפרויקט בנוגע לסעיף היכולות המקצועיות.

6.1. מעבדת דאטה לסטודנטים

הנושא של אנליטיקות מידע הוא תחום עניין מרכזי של האו"פ, והפרויקט של הקרן היה חלק משורה של פרויקטים חוצי-ארגון שעסקו באנליטיקות בכלל ובאנליטיקות למידה בפרט. החוקר הראשי של הפרויקט (פרופ' יורם קלמן) הוא היום יו"ר הועדה המוסדית לאנליטיקות למידה, שהיא גוף שמתאם את נושא האנליטיקות ברחבי האוניברסיטה ומדווח ישירות לרקטור האוניברסיטה (מכונה גם: המשנה לנשיאה לעניינים אקדמיים). בנוסף, הכניסה של אורלי וייסר לפרויקט, והקמה של ועדה אוניברסיטאית למעקב אחר הפרויקט, מאפשרת המשך פעילות מחקרית על הנושא גם אחרי סיום המימון, ובדיקת יישום רחב של אב הטיפוס על ידי האוניברסיטה בעתיד.

6.2. זיהוי ישויות בעברית

הצוות פיתח כלים ושיטות עבודה, וצבר ידע חדש במהלך העבודה בפרויקט. כמו כן פותחו יכולות של אימון רשתות נוירונים בארכיטקטורות שונות לעיבוד שפה טבעית ושל ניהול ניסויים מרובים.

6.3. עיתונות היסטורית

צוות העיתונות ההיסטורית הפך להיות מיומן בכתיבת קוד לניתוחי שפה ב-R ולוויזואליזציות שונות של ממצאי המחקר (ראו דוגמה באיור 5 בנספח).

6.4. ארכיון הרשת

במהלך השנים התגבש צוות עבודה וגוף ידע מקצועיים שאפשרו לנו להתמודד עם אתגרי חקר ארכיון האינטרנט ולהציב אותם בחזית המחקר בעולם. הצוות הוביל פרויקט ארכוב רטרוספקטיבי שדרש שילוב מיומנויות של כריית מידע, תכנות, data science וניתוח איכותני.

6.5. אודיו/וידאו

צוות האודיו גיבש במהלך השנים מיומנויות שאפשרו לנו להתמודד עם האתגרים הרבים בעיבוד מידע מגל הקול ובניתוחו. בשנה הרביעית הצוות מימש את הכישורים והידע והוביל פרויקט של ניתוח דיאלוגים השוואתי, שדרש שילוב מיומנויות של חילוץ פרמטרים אקוסטיים מאות הדיבור ושכפול המתודולוגיה על פני הקלטות מסוגים שונים, תוך אימוץ טכנולוגיות קיימות והתאמתן למקרה המבחן הספציפי.

כללי:

במסגרת הפרויקט נבנתה יכולת מקצועית של ניהול פרויקט אקדמי בינתחומי שבו חמישה חוקרים ראשיים וניהול צוות של עוזרי מחקר בהיקף בינוני. יכולת מקצועית זאת יכולה להיות לעזר בניהול מרכזי מחקר נוספים באוניברסיטה הפתוחה, דוגמת *NBEL* (המונה כ-25 סטודנטים פעילים).

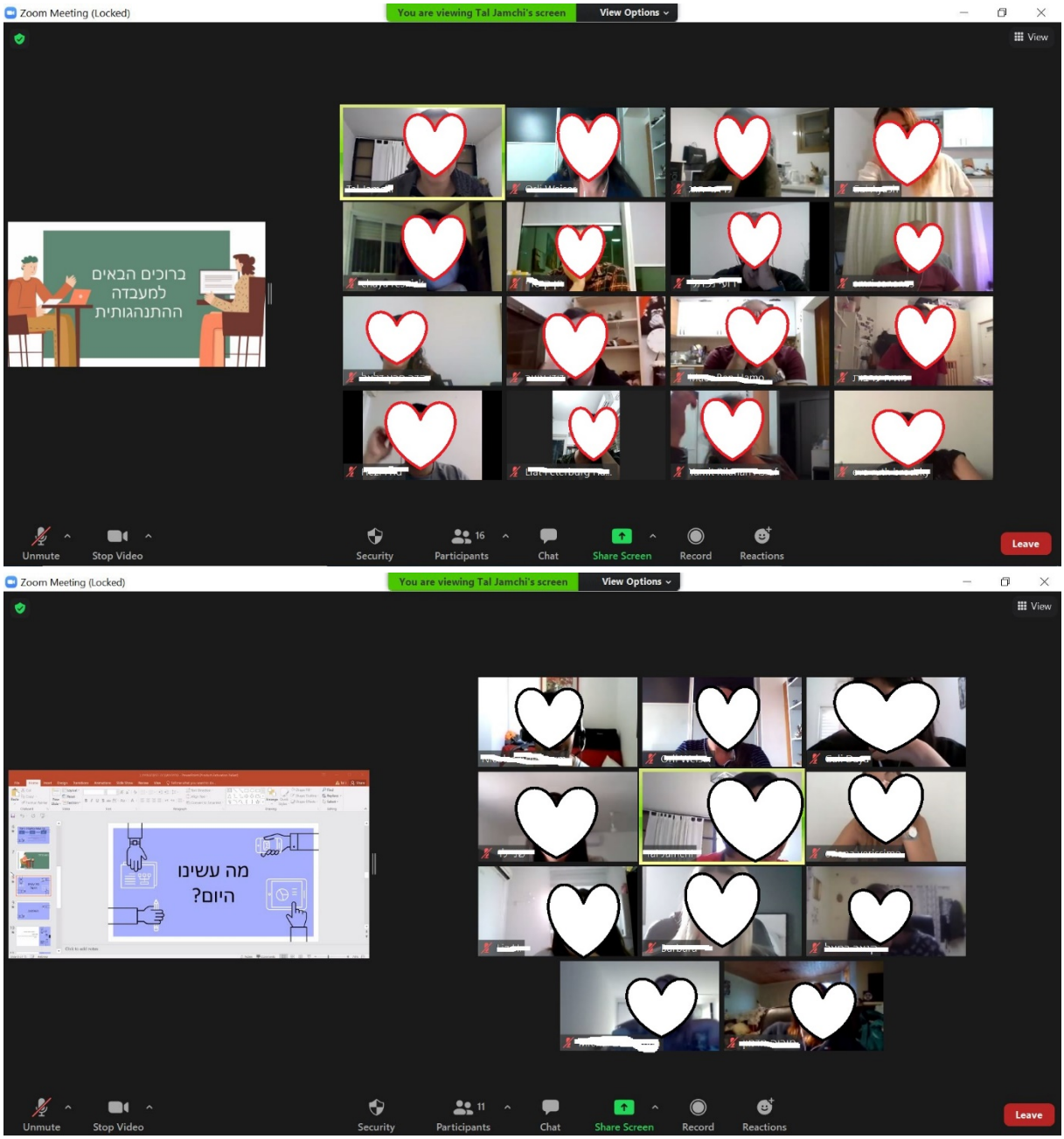
7. משוב

נשמח לקבל משוב לגבי תהליך העבודה והיחסים עם יד הנדיב וכן לקבל כל הצעה לשינוי/ לשיפור

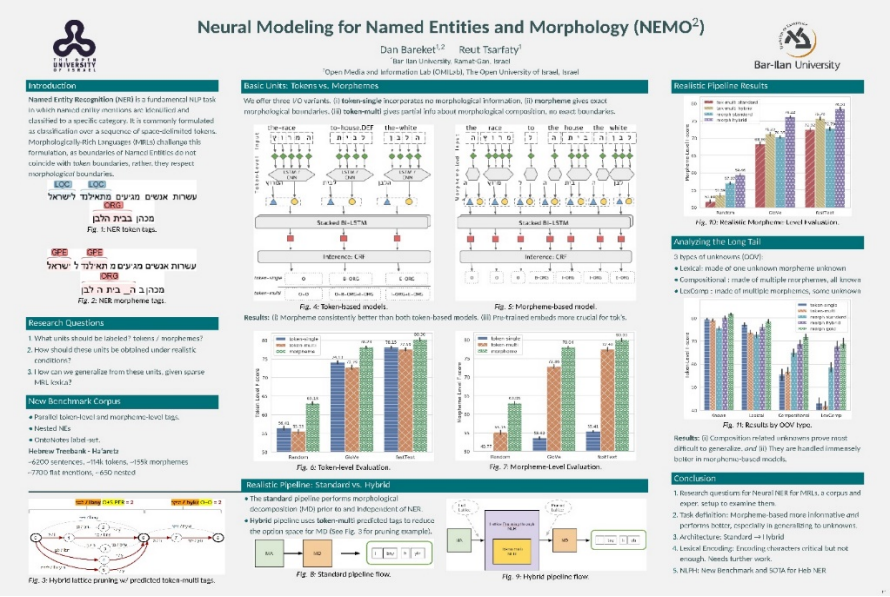
אנו מודים מאוד ליד הנדיב על המענק שסייע לקדם היבטים רבים בתחום מדעי החברה והרוח הדיגיטליים.

הדו"ח הוגש על ידי: **ד"ר ענת בן-דוד, פרופ' אורן סופר (ז"ל), ד"ר ענת לרנר, פרופ' רעות צרפתי, פרופ' יורם קלמן וד"ר ורד זילבר-ורוד**

נספח א: איורים



איור 1. צילומי מסך בזמן הפעלת הניסוי במעבדת התלמידים: בזמן תדרוך המשתתפים (למעלה) ובזמן סיכום הניסוי (למטה).



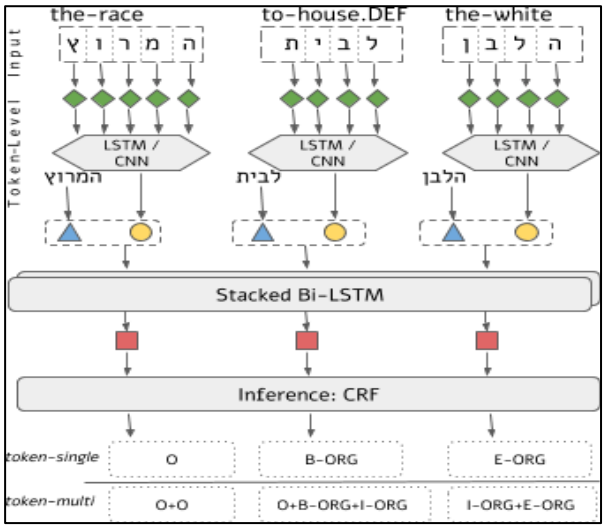


Figure 1: **The *token-single* and *token-multi* Models.**

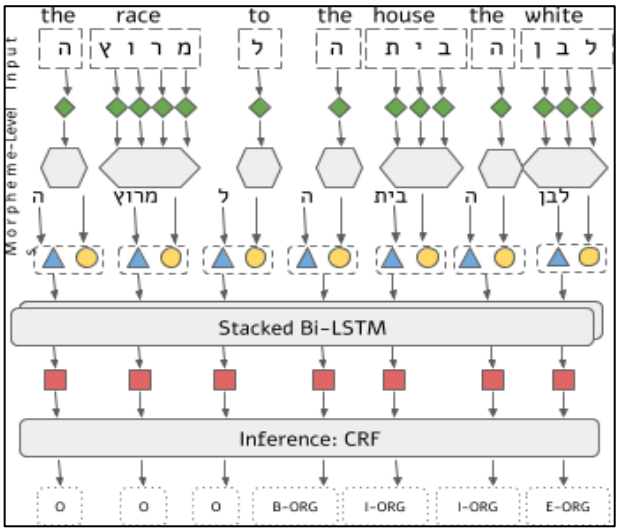


Figure 2: **The *morpheme* Model.**

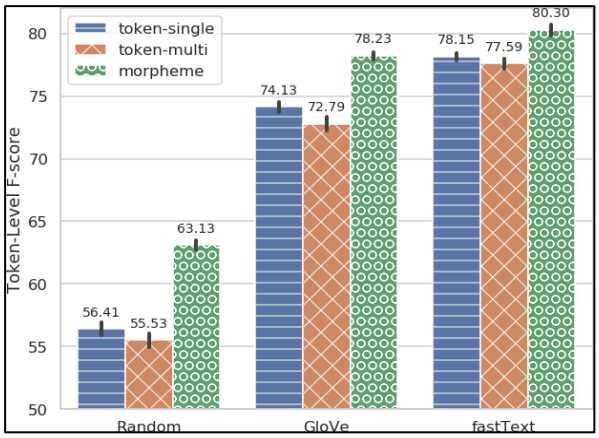


Figure 3: **Token-level Evaluation** on Dev with Gold Segmentation. Char CNN for morph, LSTM for tok.

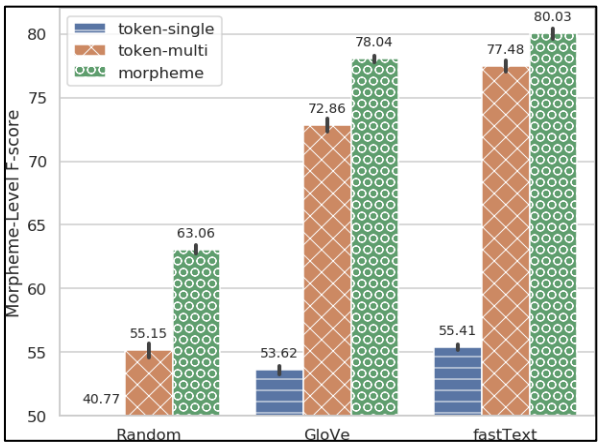
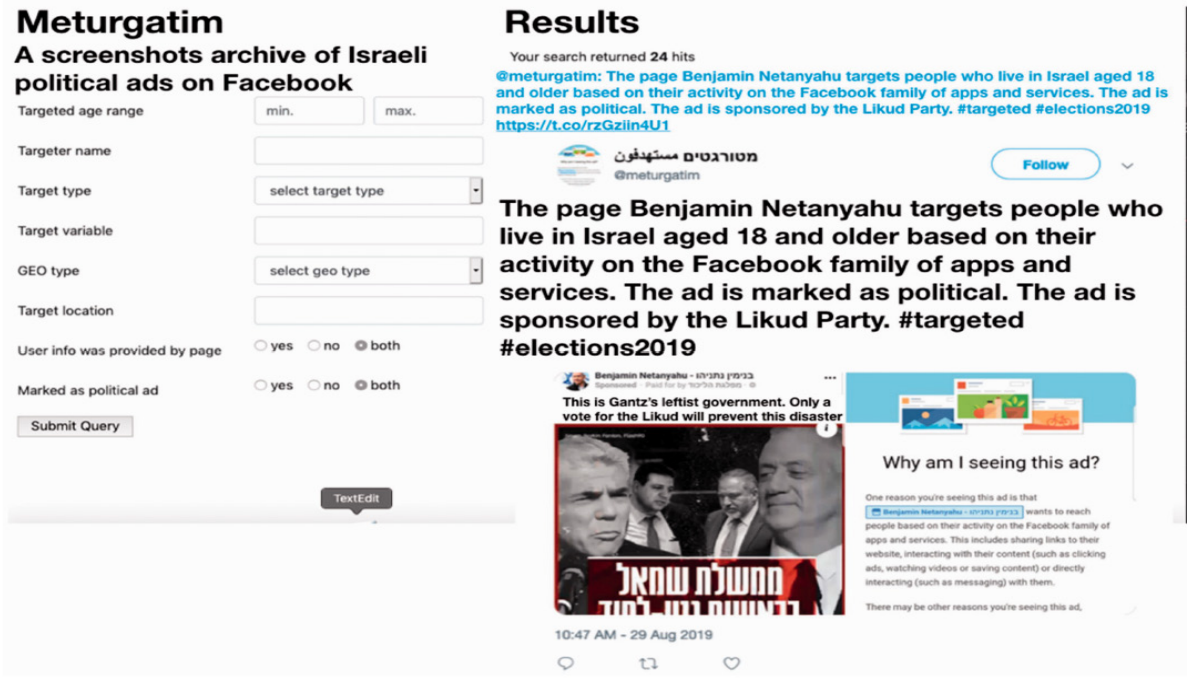
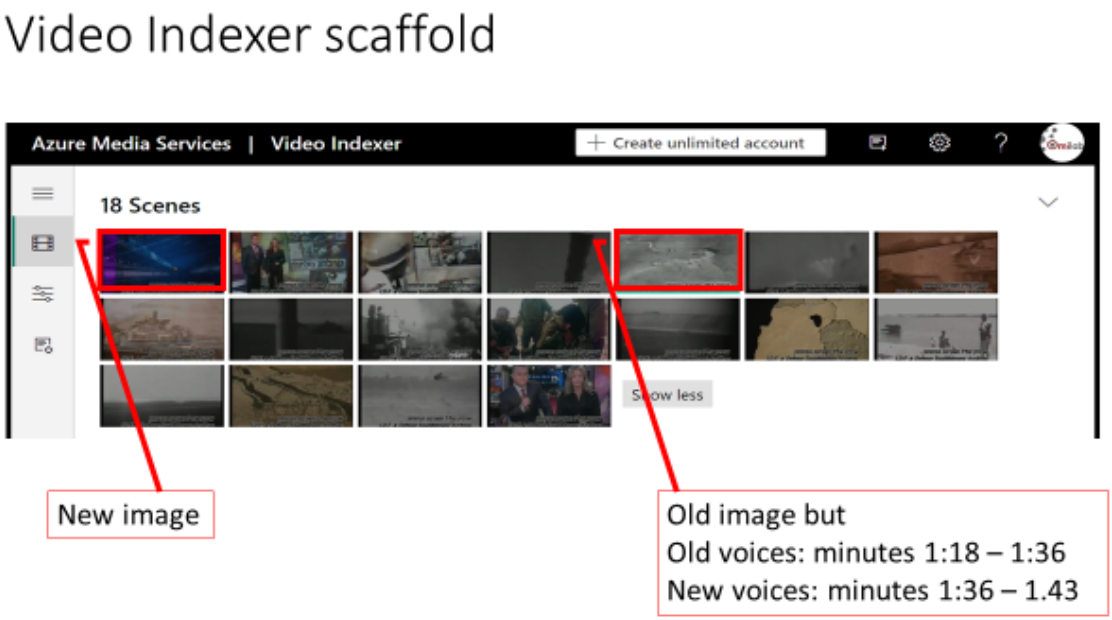


Figure 4: **Morph-Level Evaluation** on Dev with Gold Segmentation. Char CNN for morph, LSTM for tok.

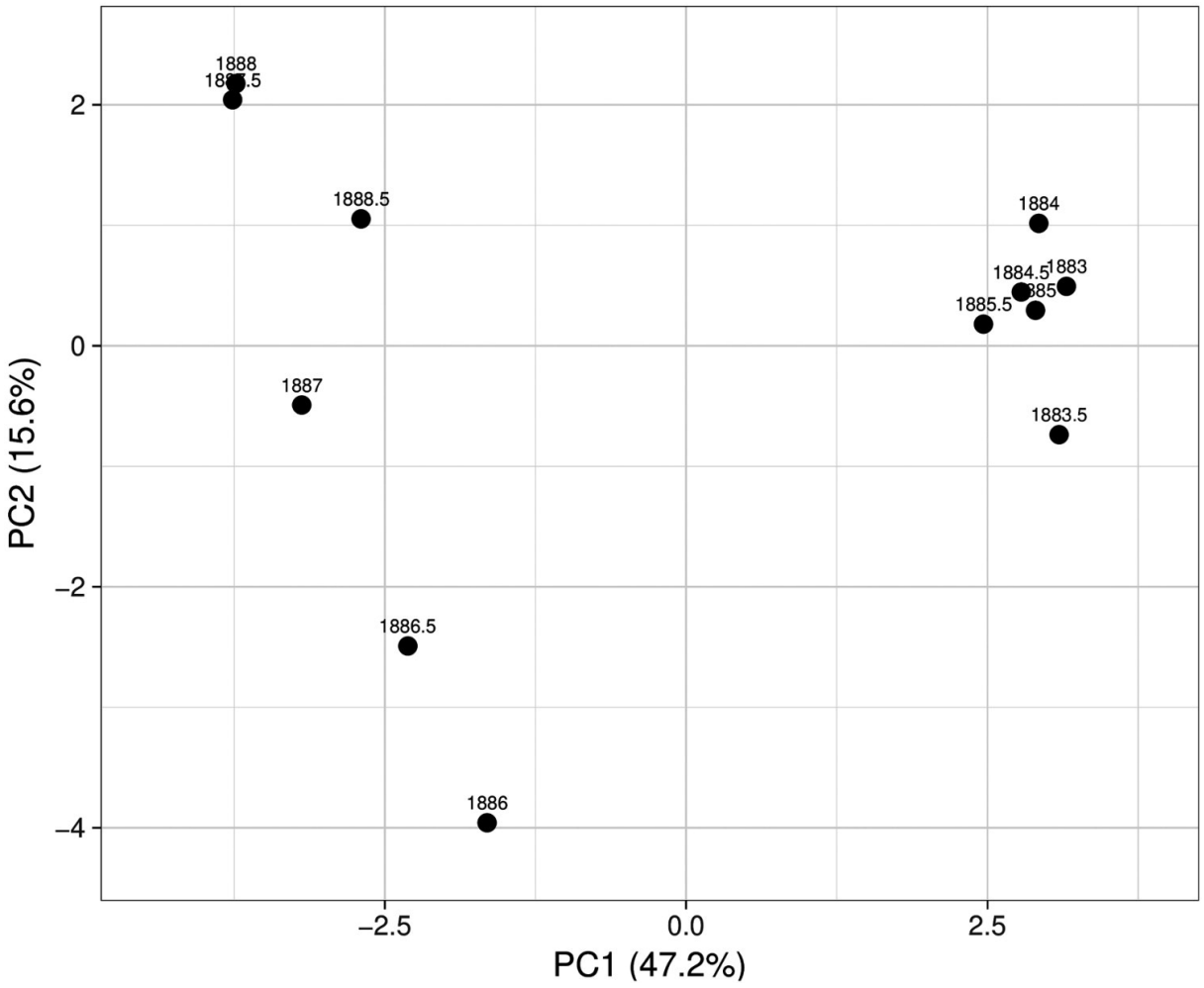
איור 2. פוסטר שהוצג בכנס ISCOL (Bareket & Tsarfaty, 2020) עם קטעים נוספים בהגדלה.



איור 3. צילום מסך מתוך ארכיון "מטורגטים" של צילומי מסך של מודעות פרסומת פוליטיות בפייסבוק. הצילום מציג את התוצאה הראשונה של "פרסומות של בנימין נתניהו שטירגטו משתמשים על בסיס הפעילות שלהם בפייסבוק. הטקסט תורגם לאנגלית מעברית על ידי ענת בן-דוד. (לקוח מתוך: Ben-David, A. (2020). (Counter-archiving Facebook. *European Journal of Communication*, 0267323120922069.



איור 4. צילום מסך של זיהוי סצנות ויזואליות של סדרה תיעודית, באמצעות פלטפורמת video indexer של מיקרוסופט. כאשר מאזינים לוידאו, הסצנה "הישנה" (מסומנת בצד ימין) היא בחלקה עם הקלטות ישנות ובחלקה עם קול של השדרן בן-הזמן. לצורך אינדוקס התאריך של הוידאו, אנו יכולים להשתמש באלגוריתם של הוידאו-אינדקסר כפיגום, אך בכדי לקבל תמונה מלאה ומדויקת, על האינדקס להשתמש בשתי האופנויות – הויזואלית והקולית.



איור 5. ניתוח גורמים ראשיים (Principal component analysis (PCA)) של תחומי שיח בשנים 1883-1888. כל ציר משקף ממד בודד של הווקטורים שזוהו על ידי האלגוריתם. הויזואליזציה נעשתה בתוכנת ClustVis (Segal & Soffer, 2020).
https://www.tandfonline.com/doi/full/10.1080/1461670X.2020.1807394

נספח ב

מכתב מפרופ' ירון צור ומר איל מילר שהתקבל ב-23 בינואר 2020:

Historical Jewish Press
עיתונות יהודית היסטורית

22.1.20

לכבוד:

פרופ' אורן סופר

ד"ר ורד זילבר-ורוד

ד"ר סיני רוזינק

ראשית כל אנו מבקשים להודות לכם על הצגת ההישגים שערכתם לנו ביום א', ולברך אתכם על התוצאות המרשימות. במאמצינו לברר את ההיתכנות של שיפור איכות ה-OCR בקורפוס העיתונות ההיסטורית עמדנו בקשר עם מספר קבוצות מחקר, שניסו מזה כחמש שנים להתמודד עם הבעיה. היו כאלו שהצליחו יותר וכאלה שהצליחו פחות, אך אין ספק שהישגיכם הניסיוניים הם המבטיחים ביותר, לפי שעה.

השלמת שלב הניסוי הראשוני שבה נמצאת שיטתכם עשוייה להפוך לנקודת מפנה משמעותית ביותר בתולדות המפגש של ישראל ומדעי היהדות עם המהפיכה הדיגיטלית. היא תציין את מועד כניסת העברית והלשונות היהודיות לרשימת הלשונות שניתן לבצע בהן מחקר דיגיטל מתקדם ללא חשש מטעויות קשות. מחקר כזה דורש לפחות 80% אחוזי דיוק, ורבים מעיתוני האתר שהודפסו באות העברית אינם עומדים בתנאי זה. המהפכה שהדיגיטציה של העיתונות היהודית יכולה לחולל בחקר מדעי היהדות לא תהיה אפוא שלמה ללא שיפור ניכר בתוצאות הOCR. שיפור כזה נראה עתה אפשרי והוא יכול להיות קפיצת המדרגה שלה כולם חיכו.

לאור זאת, אנו רואים חשיבות גדולה בבירור הסוגיות הבאות:

- איך ניתן לייצר פייפליין (PIPELINE) אוטומטי ככל הניתן ברמת כותר, כדי להריץ את המודל על כל הגליונות\שנים של כותר נתון?
- איך מוודאים שבסוף התהליך הזה התוצר המתקבל הוא תוצר פשוט ככל הניתן – המכיל את כל הרכיבים של החומר המקורי (כולל חלוקה לכתבות\מאמרים)- כדי שנוכל לעשות בו שימוש עתידי באתר פרויקט העיתונות ההיסטורית?
- מה העלויות והמשמעויות הנלוות של מהלך כזה?

אנו מצפים לתשובות מצדכם על מה שביכולתכם להשיב, כדי להבין איפה אנו עומדים מבחינת האפשרות להפוך את הישגיכם בניסויים לפרויקט מעשי ופעיל. כאמור, התרשמנו מאד מהצגת ההתקדמות ואנו מייחלים לתשובות ברורות ואמינות. ברור שאם יהיו קשיים נשב לדיון ולחשיבה משותפת. בכל מקרה, הערכתנו לעבודתכם גדולה ונשמח להמשיך לפעול ביחד.

בברכה,

פרופ' ירון צור, JPRESS מייסד ומנהל אקדמי, אוניברסיטת תל-אביב

אייל מילר, JPRESS מנהל טכנולוגי, הספרייה הלאומית