מחקר: *מחשוב ושפה*

מערכת ה- CHILDES וישומיה בחקר רכישת השפה

סיגל עוזיאל-קרל

מטרת המאמר היא לאפשר לחוקרים ולמשתמשים פוטנציאליים לערוך היכרות ראשונית עם מערכת ה-CHILDES ולהתוודע ליתרונותיו המחקריים של כלי זה. מרכיבי המערכת: תעתוק שפת ילדים, קידוד נתונים משפת ילדים, וכלים לחישוב מדדים התפתחותיים. מלבד זאת יש בה מסדי נתונים של מחקרי אורך וחתך שניתן להשתמש בהם לחקר רכישת העברית כשפת אם, וחומרי עזר נוספים.

1. הקדמה

חקר רכישת השפה מתבסס, במידה רבה, על נתונים הנאספים באמצעות תיעוד מוקלט או מצולם של אינטראקציות ספונטניות בין ילדים לבין אנשים בסביבתם הקרובה. אלה בינינו שהתנסו באיסוף נתונים מסוג זה יודעים כי בשל הקלות שניתן להפעיל בה את הרשמקול ומסרטת הווידאו מצטברות עד מהרה שעות רבות של הקלטות, אך הקלטות אלו יש לעבד, שכן איסוף הנתונים הוא רק הצעד הראשון בביצוע המחקר הבלשני. את הנתונים יש לתעתק, לקודד ולנתח. תהליך זה דורש שעות מרובות של השקעה, ולעתים אינו אמין די צורכו.

במהלך ההיסטוריה של המחקר הבלשני השתמשו חוקרי השפה בשיטות שונות לאיסוף נתונים. מקוויני (2000) מציין כי הניסיונות הראשונים להתחקות אחר אינטראקציות ספונטניות של ילדים עם הוריהם נערכו בהשראת מחקריו של דרווין על התנועות שהשתמש בהן בנו של החוקר ליצירת תקשורת. מחקרים אלה התבססו על תצפיות שתועדו בכרטיסיות או באמצעות רישום הממצאים בסוף היום, ותיארו לרוב את התפתחות השפה של ילדי החוקרים עצמם. חוקרים אלה נתקלו בבעיות מתודולוגיות שונות, כגון נתונים חסרים שלא הספיקו לרושמם בשל קצב חילופי הדברים באינטראקציה והפרעת תהליך התיעוד למהלך הטבעי של האינטראקציה עם הילד. הפרסומים של מחקרים אלה דמו מאוד לרישום

---

*ד"ר סיגל עוזיאל-קרל היא מרצה לבלשנות במחלקה לאנגלית בסמינר הקיבוצים, ובחוג להפרעות בתקשורת, אוניברסיטת חיפה. עוסקת בחקר רכישת העברית כשפת אם בגישה התפתחותית.*sigal@alum.mit.edu

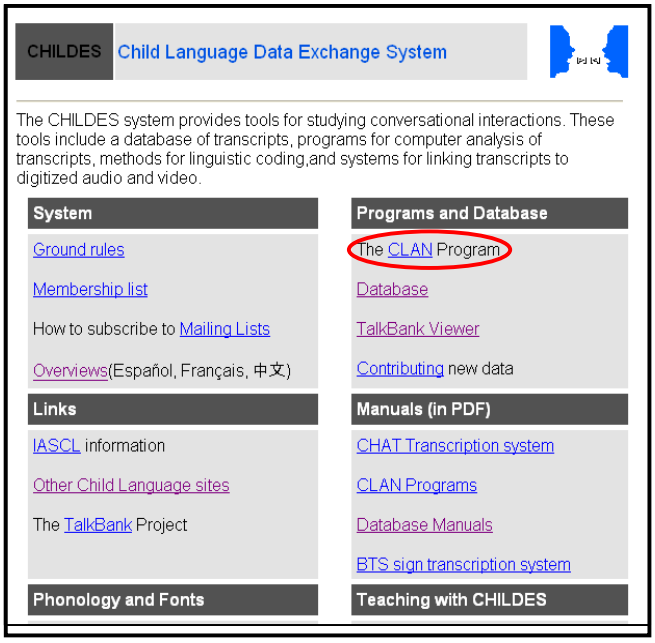
שנערך על הכרטיסיות. בסוף שנות החמישים ובשנות השישים החלו להשתמש ברשמקול לאיסוף נתונים לשוניים ואף להדפיס את תעתיקי ההקלטות שנאספו. מספרם הרב של התעתיקים בשנות השישים והשבעים לא אפשר לפרסם את כל הנתונים שנאספו במאגרי המידע, וחוקרים נאלצו לפרסם ניתוחים של נתונים אשר לא היו זמינים לחוקרים אחרים לשם אימות וביקורת.

בעשורים האחרונים, בעקבות ההתפתחות המואצת של הטכנולוגיה והאפשרות להשתמש באמצעי וידאו והקלטה משוכללים לאיסוף נתונים לשוניים, גובר בקרב חוקרי השפה הצורך בכלי ניתוח רבי-עצמה, המאפשרים עיבוד כמויות גדולות של נתונים לשוניים במגוון רחב של אופנים, במהירות וברמת דיוק גבוהה, מתוך רצון לשפר את איכות המחקר הבלשני. זמינותו של המחשב, היותו חלק כמעט בלתי-נפרד מחיינו והיכולות הטכנולוגיות הגלומות בו, הפכו אותו לאמצעי המועדף לפיתוח כלים מתקדמים לעיבוד נתונים לשוניים.

מתוך תפיסה זו פותחה בתחילת שנות השמונים מערכת ה- CHILDES (**Chi**ld **L**anguage **D**ata **E**xchange **S**ystem). זו מערכת ממוחשבת הכוללת ממשק לתעתוק, קידוד וניתוח של נתונים לשוניים משפות שונות באמצעות כלים ייעודיים מגוונים ופשוטים לשימוש (MacWhinney, 2000). כפי שמרמז שמה, מערכת ה- CHILDES פותחה לצורך התמודדות עם ניתוח שפת ילדים דבורה, ובעיקר עם היבטים שונים של רכישת השפה, כגון פונולוגיה, מורפולוגיה ותחביר. אולם באותה מידה יכולה המערכת לשמש לניתוח שלבים שונים של רכישת שפה מאוחרת ואף לניתוח שפת מבוגרים בתרחישים שונים (שיחה, טקסטים מונולוגים) ובאופנויות שונות (שפה דבורה וכתובה).

לעיבוד נתונים ממוחשב ואוטומטי יתרונות אחדים. ראשית, ניתן לנתח באופן זה ***כמויות גדולות*** של נתונים משפות שונות, ממספר רב של נבדקים ומרמות גיל שונות. שנית, ניתן לבצע מגוון רחב של ניתוחים מסוגים שונים באופן ***מהיר, מדויק וזול*** יחסית. בנוסף, היות שמערכת ה-CHILDES מאפשרת ניתוחים של קובצי תעתיק הכתובים על פי כללים מסוימים, הדבר מחייב הקפדה על ***אחידות ועקביות*** בתעתוק ובקידוד, ובאופן זה יוצר שקיפות בתהליך המחקרי. הכללים האחידים להזנת הנתונים, ללא תלות בשפה הנחקרת, והיותה של המערכת ממוחשבת וזמינה דרך רשת האינטרנט, מאפשרים ***שיתוף פעולה מחקרי***, כמו גם שחזור ממצאים של מחקרי עמיתים ו***ביקורת***, ואלה תורמים בסופו של דבר לשיפור איכותו של המחקר הבלשני.

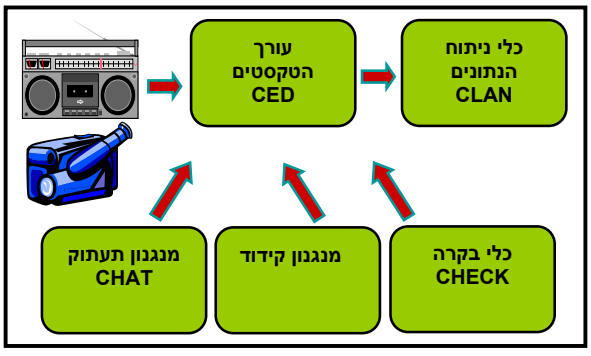
מאמר זה מציג את מערכת ה- CHILDES על מרכיביה השונים, תוך התמקדות ביישומיה לחקר רכישת שפה בכלל ולחקר העברית כשפת אם בפרט. תחילה תוצג המערכת באופן כללי, ואחר כך ייסקרו מרכיבים מספר ביתר פירוט: תעתוק שפת ילדים, קידוד נתונים משפת ילדים וכלים לחישוב מדדים התפתחותיים. בנוסף יסקרו מדורים באתר ה-CHILDES שבהם יש חומרי עזר רלוונטיים בתחום רכישת השפה – מסדי נתונים של מחקרי אורך וחתך שניתן להשתמש בהם לחקר רכישת העברית כשפת אם, קישורים לאתרים נוספים העוסקים ברכישת שפה, הביבליוגרפיה של CHILDES בתחום רכישת השפה, וקבוצת הדיון שבה נדונות שאילתות מתודולוגיות הקשורות לתכנה ויישומיה מחד גיסא, ושאילתות תאורטיות הנוגעות לרכישת שפה מאידך גיסא. מטרת המאמר היא לאפשר לחוקרים המעוניינים בכך ולמשתמשים פוטנציאליים לערוך היכרות ראשונית עם המערכת ולהתוודע אל יתרונותיו המחקריים של כלי זה.



תרשים 1: דף הבית של אתר CHILDES

2. סקירה כללית של מערכת ה-CHILDES

מערכת ה-CHILDES היא, כאמור, מערכת ממוחשבת לתעתוק, לקידוד ולניתוח נתונים לשוניים. תוכנת ה-CHILDES ניתנת להורדה ללא תשלום מאתר המערכת (http://childes.psy.cmu.edu), באמצעות לחיצה על קישור (The CLAN Program) שנמצא בדף הבית תחת הכותרת Programs and Databases (ראו תרשים 1) ומילוי ההנחיות שם.



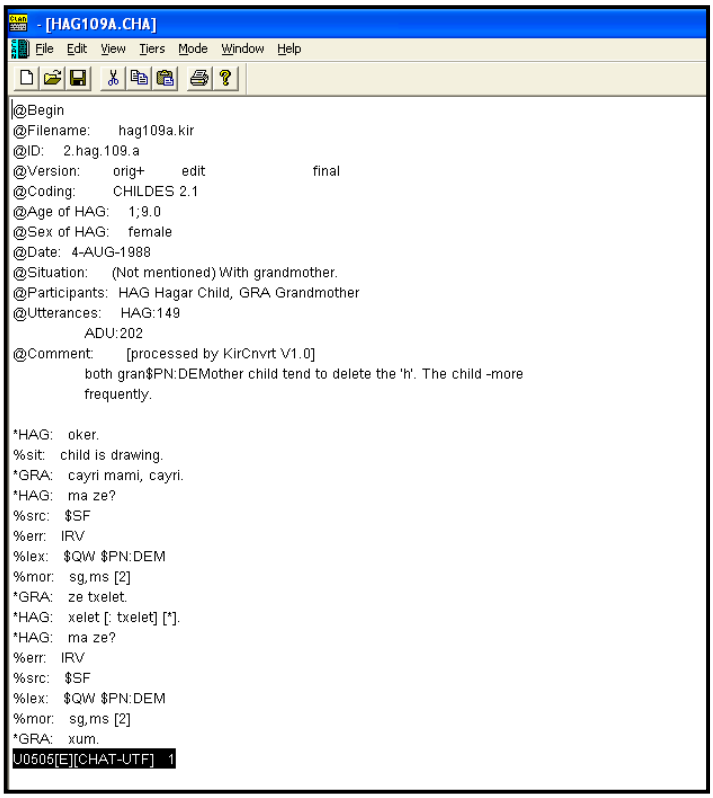
תרשים 2: מבנה מערכת ה-CHILDES

למערכת מרכיבים מספר: ***עורך הטקסטים, מנגנוני התעתוק והקידוד, כלי הבקרה וכלי הניתוח***. מבנה

המערכת מתואר באופן סכמטי בתרשים 2.

2.1 עורך הטקסטים

לאחר ההתקנה של תוכנת ה-CHILDES מופיעה על שולחן העבודה צלמית של התכנה. לחיצה על הצלמית פותחת קובץ חדש. אליו ניתן לתעתק נתונים לשוניים על-פי כללי התעתיק הנהוגים בעברית וכן על-פי הפורמט של מערכת ה- CHILDES.

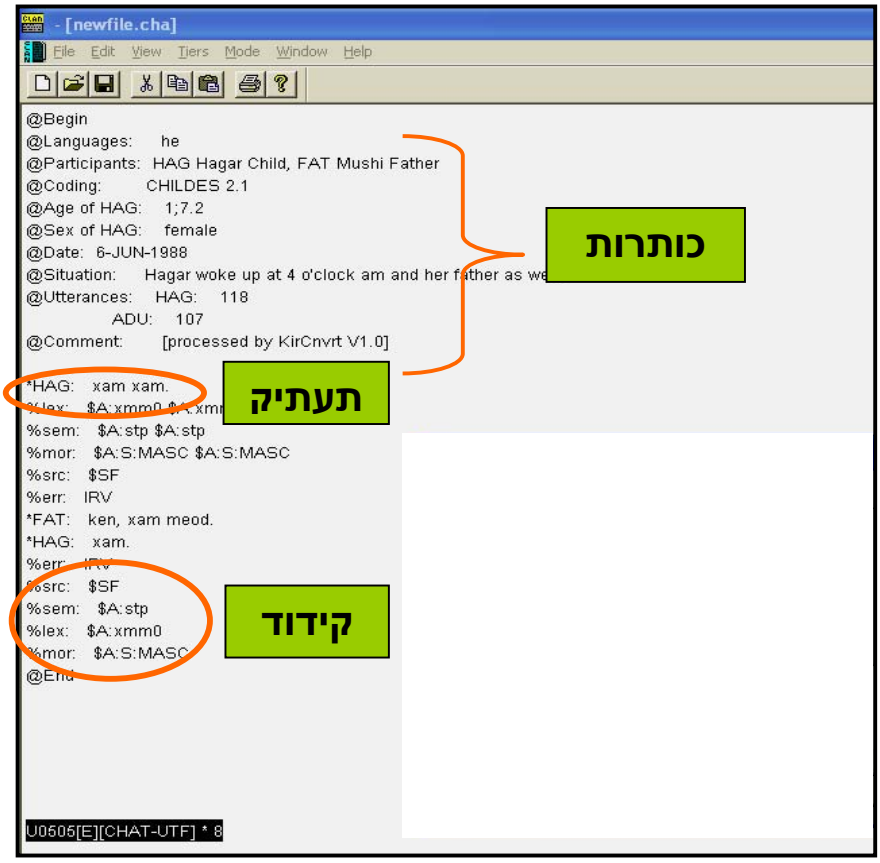


תרשים 3: חלון העריכה

כפי שניתן לראות בתרשים 3, קובץ ה-CHILDES דומה במבנהו לקובץ סטנדרטי של מסמכי WORD. הוא כולל סרגל כלים. מקצת כליו זהים לכלים של ה-WORD (למשל, קובץ, עריכה, עזרה) ומקצתם ייעודיים לתוכנה זו. כברירת מחדל נפתח קובץ חדש במצב של עריכה. מצב 'עריכה' מאפשר לבנות קובצי תעתיק חדשים או לערוך קבצים קיימים. במהלך התעתוק או בסיומו ניתן לבדוק את מבנה הקובץ כדי לוודא שהוא תואם את כללי המערכת באמצעות הרצת תכנית בדיקה הנקראת CHECK בתוך חלון פקודות מיוחד שנועד לצורך זה.

2.1.1 מבנה קובץ התעתיק

לקובץ התעתיק שלושה חלקים עיקריים: ***כותרות***, ***שורות תעתיק*** ו***שורות קידוד*** (ראו תרשים 4), והוא בעל מבנה קבוע וקשיח הנקבע על פי כללים מוסכמים. כך ניתן לשמור על אחידות בתעתיק בשפות שונות.



תרשים 4: קובץ התעתיק

קובץ תעתיק חדש יש לשמור בסיומת "cha", כדי שהמערכת תזה אותו. כל קובץ נפתח במילה @Begin ומסתיים במילה @End. בראש הקובץ מופיעות כותרות, מקצתן כותרות חובה, ומקצתן כותרות רשות שהמתעתק יכול להוסיפן על-פי צרכיו. כל כותרת מסומנת בתחילתה באמצעות הסימן @. מתחת לכותרות מופיעות שורות התעתיק. התעתיק נכתב באותיות לטיניות על-פי מוסכמות תעתיק שנקבעו והותאמו לעברית (ברמן, 1989). כל שורת תעתיק נפתחת בציון שם הדובר בשלוש אותיות גדולות ולפניהן הסימן *. לדוגמה, הצירוף *MOT: מציין כי הדוברת היא אמה של סמדר (**Mot**her). לאחר שם הדובר מופיעה שורת התעתיק. בסופה חייב להופיע אחד מסימני הפיסוק: נקודה, סימן שאלה או סימן קריאה. בנוסף לסימני הפיסוק הללו, ניתן להשתמש בסימנים המציינים חריגות שונות בשטף הדיבור. להלן מספר דוגמות רלוונטיות לתעתוק שפת ילדים:

- כאשר מילים בתמליל המתועתק אינן ברורות, הן יסומנו ברצף של שלושה Xים:

*CHI: yesh li xxx sham.

- כאשר מילה בתמליל מופיעה באופן חלקי, אך ניתן לשחזר את המשכה בבירור, יתועתק ההמשך בתוך סוגריים עגולים:

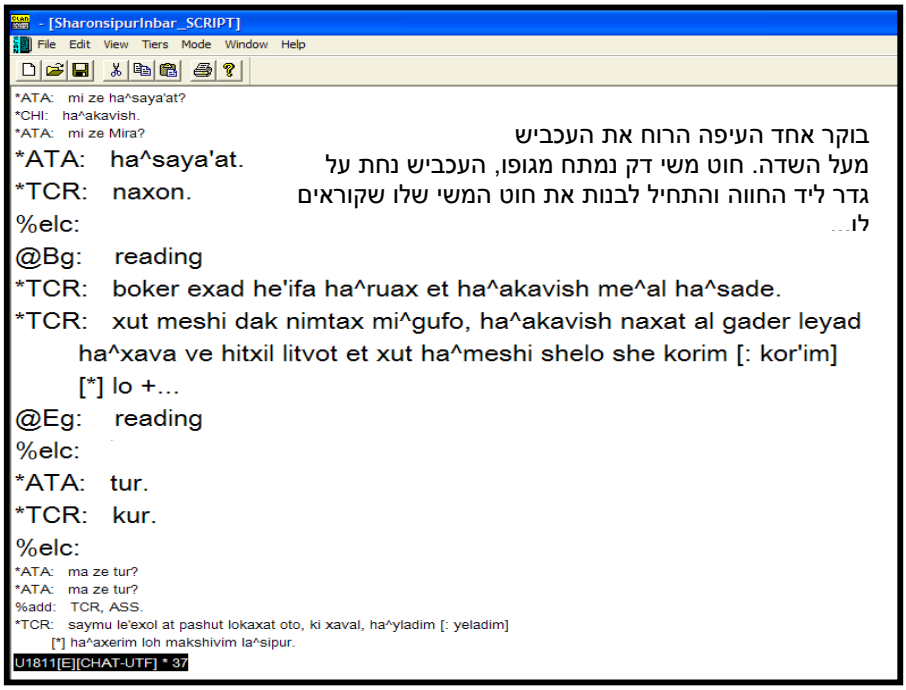
*CHI: yashav(ti) kol ha^yom.

- שגיאה תסומן בתוך התעתיק כ- [*][: text]:

*CHI: ofo [: eyfo][*] aba?

- הפסקה קצרה בין הברות מסומנת ב # והפסקה ממושכת ב ##:

*MOT: ze hipo#potam?

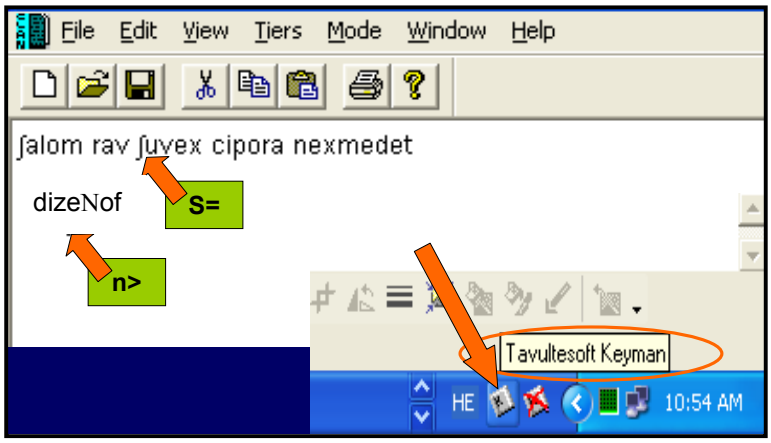


תרשים 5: דוגמא לתיחום קטע מיוחד

תכנת ה-CHILDES מאפשרת לסמן קטע מסוים כקטע מיוחד על ידי תיחומו באמצעות הסימנים @Bg בתחילתו ו– @Eg בסופו, כמתואר בתרשים 5. בתרשים נתחם קטע הסיפור על העכביש וסומן כקטע קריאה. תיחום זה מאפשר לחוקר לנתח קטעים אלה ניתוחים לשוניים תוך התעלמות משאר התעתיק, או, לחילופין, לנתח את התעתיק ללא התייחסות לקטעים המיוחדים.הקפדה על בניית קובץ התעתיק על-פי הכללים מונעת תקלות בהרצת פקודות הניתוח בשלבים הבאים.

2.1.2 תעתוק הנתונים

תוכנת ה-CHILDES תומכת בשתי צורות תעתיק: תעתיק פונמי (רחב) ותעתיק פונטי (צר). שני סוגי התעתיק נכתבים, כאמור, באותיות לטיניות על-פי כללי תעתיק שהותאמו לעברית. התעתיק הפונמי, הרחב, אינו מייצג במדויק שגיאות הגייה, הבדלים בין דיאלקטים, הטעמה וכדומה. לעומת זאת, התעתיק הפונטי, הצר, מאפשר תעתוק מפורט ומדויק של הנתונים והוא מתאים לתיעוד רכישת הפונולוגיה ולקויות שפה הקשורות בהגייה. השימוש בתעתיק הפונטי ב-CHILDES מצריך התקנה של תכנה נוספת הניתנת להורדה ללא תשלום מאתר המערכת, ומחייב שימוש בטבלאות המרה. בטבלאות אלה מופיעות רשימות של צירופי אותיות וסימנים. אלו מיתרגמים על ידי התוכנה הנוספת לסימנים פונטיים. דוגמה לשימוש בתעתיק זה מובאת בתרשים 6.



תרשים 6: תעתיק פונטי

בתחתית התרשים ניתן לראות את הצלמית של התוכנה הפונטית. כשהצלמית מופיעה ללא ה-X האדום (השוו לצלמית הימנית), ניתן להשתמש בכתיב הפונטי על ידי הקלדת צירופי האותיות הרצויים על-פי טבלאות ההמרה. לדוגמה, בשורה העליונה בתרשים מופיע המשפט 'שלום רב שובך, ציפורה נחמדת'. כדי לקבל את הסימן הפונטי ל'ש', למשל, יש להקליד את האות S ולאחריה =. בשורה התחתונה בתרשים מתועתקת המילה 'דיזנגוף'. כדי לתעתק את הצירוף 'נג' בכתיב הפונטי, יש להשתמש באות N ולאחריה הסימן <. השימוש בטבלאות ההמרה מהווה חיסרון בשל היותו של התעתיק בלתי רציף ולעתים אף מייגע.

תוכנת ה-CHILDES מאפשרת לסמן מילים בשורת התעתיק כשייכות לקבוצה מסוימת. באופן זה ניתן לאתרן בקלות בתעתיקים, ובמידת הצורך לא להתחשב בהן בניתוחים הלשוניים. דוגמאות לקבוצות מילים כאלה מובאות בטבלה 1.

קטגוריות	קודים	דוגמה	משמעות
מלמול	@b	abame@b	—
מילה מומצאת	@c	mangid@c	פועל כלשהו
צורה ייחודית למשפחה	@f	nicanon@f	שם חיבה לניצן (ילד)

קטגוריות	קודים	דוגמה	משמעות
תחדיש	@n	nifraq@n	התפרק
אונומטופיאה	@o	ananan@o	מכונית נוסעת
מילות שיר	@si	lalala@si	שיר
פריט מבחן	@t	wug@t	יצור קטן
משחק מילים	@wp	talgal@wp	טל-גל (חריזה)

טבלה 1: סימון קבוצות מילים מיוחדות

2.2 קידוד הנתונים

עורך הטקסטים משמש לקידוד התעתיקים כשהוא נמצא במצב 'קידוד'. ניתן לעבור ממצב עריכה למצב קידוד באמצעות פתיחת תפריט 'מצב' (Mode) בסרגל הכלים בקובץ התעתיק ובחירת 'מצב קידוד' (Coder Mode). כמערכת קידוד מספקת מערכת ה-CHILDES למקודד דרך שיטתית ועקבית לקידוד תעתיקים מתוך תפריטי קטגוריות מוגדרים מראש. שורת הקידוד מופיעה תמיד מתחת לשורת התעתיק המקודדת והיא מסומנת ב- % ולאחריו שלוש אותיות המציינות את קטגוריית הקידוד. טבלה 2 מציגה דוגמה לקידוד הפועל 'נפל' בקטגוריות אחדות.

קידוד	קטגוריית הקידוד	דוגמה - *nafal*	*CHI:
לקסיקלי	%lex	V:*npl*1	%lex:
מורפולוגי	%mor		
סמנטי	%sem		%mor: V:S:MASC:3:PAST

תחבירי %syn

%sem: V:sch (change of state)

%syn: VP

טבלה 2: קידוד הנתונים ב-CHILDES

בדוגמה שבטבלה המבע 'נפל' מקודד קידוד לקסיקלי (פועל, שורש ובניין), קידוד מורפולוגי (מין, מספר, גוף וזמן), קידוד סמנטי וקידוד תחבירי. הקידוד הלקסיקלי מיוצג על ידי lex%. בשורת קידוד זו, V מייצג את קטגוריית הפועל, האותיות "npl" מציינות את השורש ואילו הספרה "1" מייצגת את הבניין (קל). הקידוד המורפולוגי מיוצג על ידי mor%. גם כאן מצוינת קטגורית הפועל באמצעות האות V, לאחריה האות S המייצגת את צורת היחיד, הצירוף MASC המציין "זכר", הספרה "3" המציינת גוף שלישי ולבסוף המילה PAST המציינת זמן עבר. הקידוד הסמנטי מיוצג על ידי sem%. הוא כולל את האות V לציון הפועל, ולאחריו רצף האותיות "sch" שמשמעו שינוי מצב (**ch**ange-of-**s**tate). הקטגוריה התחבירית מיוצגת על ידי syn% והיא כוללת את הצירוף VP שמשמעו 'צירוף פעלי'.

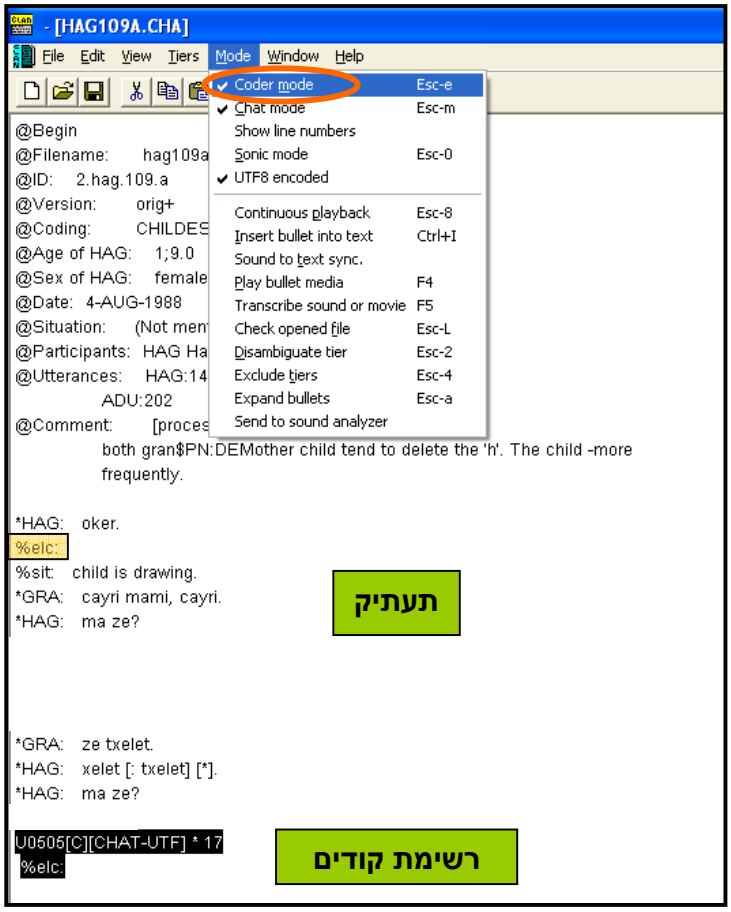
המערכת מאפשרת קידוד נתונים באחד משלושה אופנים: (1) ידני , (2) חצי-אוטומטי ו-(3) אוטומטי.

2.2.1 קידוד ידני

הקידוד הידני דורש הקלדה של שורות הקוד בעבור כל מבע מחדש. צורת קידוד זו מתאימה למסד נתונים קטן, מאחר שהיא מייגעת ועשויה לפגום בעקביות הקידוד.

2.2.2 קידוד חצי-אוטומטי

הקידוד החצי-אוטומטי מבוסס על בחירת קודים מתוך רשימה מוכנה מראש. רשימות קודים מסוימות מופיעות באתר המערכת ואין צורך לבנותן (למשל, רשימת קודים לניתוח השיח). לצורך ביצוע קידוד חצי-אוטומטי יש לפתוח את קובץ התעתיק במצב 'קידוד' (ראו תרשים 7). במצב זה, המסך מתפצל לשני חלקים: בחלק העליון מופיע התעתיק שרוצים לקודד ובתחתון רשימת הקודים. כדי לקודד מבע כלשהו, יש למקם את סמן העכבר על שורת התעתיק הרצויה ולבחור מתוך הרשימה את הקוד המתאים.



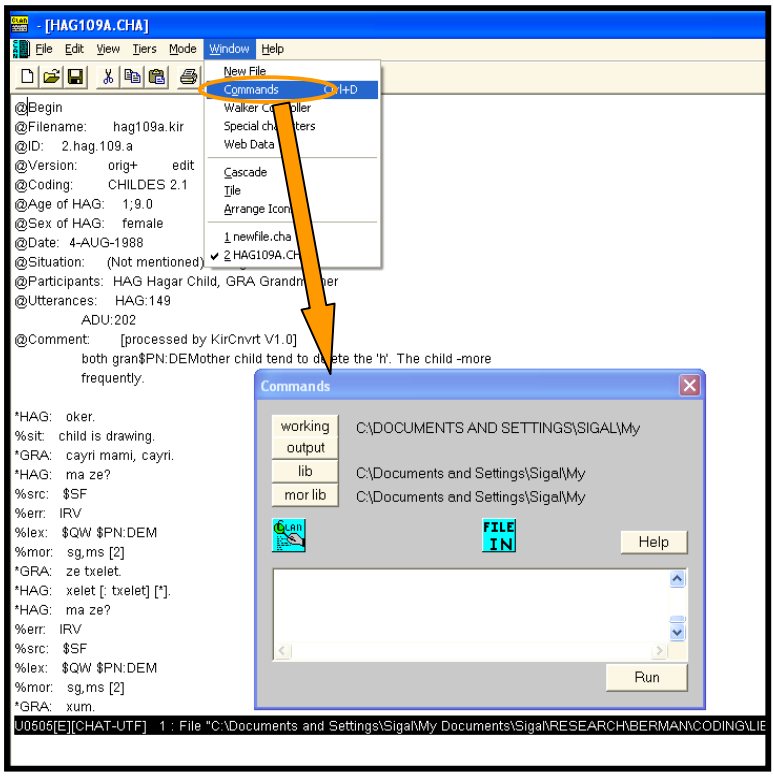
תרשים 7: קידוד חצי-אוטומטי

כאשר לוחצים על מקש ה-ENTER הקוד מופיע בשורת הקידוד מתחת לשורת התעתיק המתאימה (ראו דוגמה). ניתן לחזור על תהליך זה מספר פעמים רב ולקודד באופן זה מספר רב של מבעים בתעתיקים שונים וברמות קידוד שונות.

2.2.3 קידוד אוטומטי

במקרים מסוימים מערכת ה-CHILDES מאפשרת קידוד נתונים אוטומטי לחלוטין. כדי להפעיל את הקידוד האוטומטי יש לפתוח חלון פקודות מיוחד ובו לכתוב את פקודת הקידוד. את חלון הפקודות ניתן לפתוח באמצעות בחירת 'חלון' בסרגל הכלים של קובץ התעתיק, ובתפריט שנפתח לבחור ב'פקודות' (ראו תרשים 8). לאחר שחלון הפקודות נפתח יש לרשום את פקודת הקידוד וללחוץ על ENTER. הרצת הפקודה מאפשרת קידוד שיטתי של מספר קבצים רב בו-זמנית, ללא כניסה לכל קובץ תעתיק בנפרד.

תעתיקים של עברית ניתן לקודד באופן אוטומטי רק ברמה המורפולוגית, באמצעות כלי שפותח לצורך זה על ידי המחברת ועל ידי ברכה ניר-שגיב מאוניברסיטת תל אביב http://childes.psy.cmu.edu/(morgrams/).. המקודד המורפולוגי של העברית כולל כ- 30,000 ערכים. הערכים נבחרו על פי שכיחותם בקורפוס מחקר האורך של ברמן (1991). אף אותו ניתן להוריד מאתר המערכת: http://childes.psy.cmu.edu/manuals/12other.doc#_Ref503958473.



תרשים 8: חלון הפקודות

המקודד המורפולוגי של העברית הוא בעל יכולת ניתוח רחבה: הוא מאפשר ניתוח מורפולוגי של שמות עצם, תארים, פעלים ומילות יחס. הוא כולל מילונים של מילים בלתי מנותחות כמו תואר הפועל (במהירות, באמת), שמות פרטיים (סמדר, ליאור), שמות מספר, מילים בשפת הילדים (לוצה = לא רוצה, ביצות) ומילים אונומטופאיות (הב הב, קוקוריקו). כלי זה מאפשר ניתוח של מילים במשלבים שונים (איפה –היכן,למה –מדוע), ומציע מספר אפשרויות ניתוח למילים הומונימיות (בה, בא, ב-).

2.3 ניתוח הנתונים

בתום תהליך התעתוק והקידוד מתבצע תהליך ניתוח הנתונים. כלי הניתוח של ה-CHILDES הם אוטומטיים, וניתן להפעילם על מספר גדול של קבצים בו-זמנית באמצעות הרצת פקודה בודדת בחלון הפקודות, כפי שתואר לעיל. כלים אלה מגוונים ומותאמים לניתוח כמותי ואיכותי של נתונים לשוניים שתועתקו וקודדו בפורמט ה-CHILDES. כלי הניתוח של המערכת מאפשרים ביצוע ניתוחים לשוניים מסוגים שונים על תעתיקים בגדלים שונים. הניתוחים מהירים ומדויקים ותוצאותיהם ברורות למקרא.

מערכת ה-CHILDES מאפשרת ביצוע ניתוחים על קובץ יחיד או על קבצים מספר לפי חתכים שונים, כגון: דובר מסוים, קטגורית קידוד מסוימת או קוד מסוים, וכן ניתוחים המצליבים בין קטגוריות קידוד שונות. לדוגמה, ניתן למצוא כמה פעמים משתמש נבדק מסוים בפועל כלשהו בזמן עבר. במקרה זה מוצלב מידע לקסיקלי על הפועל (שורש + בנין) עם מידע מורפולוגי (זמן דקדוקי).

בין מגוון הניתוחים שהמערכת יכולה לבצע נכללים: חיפושי מילים או מחרוזות מילים, ניתוחי שכיחויות וחישוב מדדים להתפתחות לשונית (גיל לשוני, אורך מבע ממוצע, אורך תור ממוצע, גיוון לקסיקלי). מדדים אלה ורבים נוספים מתוארים בפירוט באתר ה- CHILDES (http://childes.psy.cmu.edu/manuals/CLAN.pdf).

2.3.1 חיפושי מילים ומחרוזות מילים

באמצעות פקודות החיפוש ב-CHILDES ניתן לאתר מילה או מחרוזת מילים בנתונים של דובר אחד או קבוצת דוברים. למשל, ניתן ליצור רשימה של כל המבעים בתעתיק מסוים אשר מופיעה בהם מילה או צירוף מילים כלשהו (למשל, 'רוצה', 'לא', או 'לא + רוצה') [KWAL]. באמצעות פקודות החיפוש ניתן לאתר מילה בעלת קוד מסוים (למשל, כל המבעים שיש בהם פועל)

[COMBO], ואף רשימה של כל המבעים שהטקסט בהם ייחודי (למשל, שורות מתוך סיפור)

[GEMLIST].

2.3.2 ניתוחי שכיחויות

באמצעות פקודות הניתוח ב-CHILDES ניתן לבצע חישובי שכיחויות מסוגים שונים על חתכים שונים של מסד הנתונים – על שורות התעתיק או על שורות הקידוד, על נתונים של דובר בודד או של קבוצת דוברים. ניתן, למשל, לחשב מהי שכיחות הופעתן של מילים מסוימות במסד הנתונים [FREQ], או מהי שכיחות הופעתן של מילים מסוימות יחד (למשל, כמה פעמים מופיעה המילה 'מה' עם המילה 'קרה') [COOCCUR] . ניתן לחשב מהי שכיחות הופעתם של חלקי דיבר שונים, או מהי שכיחות הופעתם במסד הנתונים של מבנים תחביריים שונים כגון, משפטי שאלה.

2.3.3 חישובי מדדים להתפתחות לשונית

מערכת ה-CHILDES פותחה מלכתחילה לעיבוד נתונים לשוניים משפת ילדים. לאור זאת, אחד המאפיינים הייחודיים שלה הוא מגוון המדדים שהיא מציעה לקביעת התפתחות לשונית. מדדים אלה נחשבים התפתחותיים, מאחר שהערך המספרי שמתקבל מהם עולה ככל שהילד מפותח יותר מבחינה לשונית. בין המדדים הללו נכללים: **אורך מבע ממוצע** (Mean Length of Utterance – MLU) – מתאר כיצד גדל מספר המורפמות (או המילים) של הילד בכל מבע ככל שהוא מתפתח מבחינה לשונית (דוגמה לחישוב מדד זה מובאת להלן). **אורך תור ממוצע** (Mean Length of Trun – MLT) – מתאר כיצד גדל מספר המבעים של הילד בכל תור ("תור" מוגדר **כרצף** מבעים של דובר מסוים). **יחס תבנית-תמנית** (Type-Token Ratio –TTR) – מתאר את היחס בין מספר המילים השונות של הדובר בתעתיק מסוים (תבניות) למספר הפעמים שבהן הוא משתמש בכל מילה (תמניות), **וגיוון לקסיקלי** (Vocabulary Diversity – VOCD) – מנתח את ההסתברות שאלמנטים לקסיקליים חדשים יופיעו בתעתיקים ככל שאורכם גדל.

חישוב **אורך מבע ממוצע** יודגם באמצעות האינטראקציה הבאה בין אם לבין בנה:

אם: מה זה?

בן: ספר.

אם: מה יש פה?

אם: מי זה?

בן: ילד.

אם: מה הילד עשה?

אם: מה הילד עשה?

בן: אכל.

בדוגמה זו מניין המבעים של האם הוא חמישה ושל הילד שלושה. מספר המורפמות הכולל במבעי הילד הוא 3, ואילו במבעי האם – 15 (לצורך הדוגמה, הא הידיעה נספרה כמורפמה נפרדת, והפועל בגוף שלישי יחיד בעבר נספר כמורפמה אחת). חישוב **אורך מבע ממוצע** נעשה על ידי חילוק מספר המורפמות הכולל של דובר מסוים במספר המבעים הכולל של אותו דובר בתעתיק נתון (ניתן לחשב אורך מבע ממוצע גם במילים). בדוגמה לעיל מספר המבעים של האם הוא 5 ומספר המורפמות 15. מנת החילוק של 15 ב-3 היא 5. לפיכך, אורך המבע הממוצע של האם הוא 5. לעומת זאת, אורך המבע הממוצע של הבן הוא 1 (3:3=1).

MLU -f +t*MOT sample.cha

Sat Jun 19 12:16:15 2004mlu (04-Dec-2003) is conducting analyses ONLY on speaker main tiers matching: *MOT

****

From file <sample.cha>

MLU for Speaker: *MOT:

Number of: utterances = 3, morphemes = 15

Ratio of morphemes over utterances = 5.000

Standard deviation = 0.816

הדוגמה לעיל מציגה פלט של חישוב אורך מבע ממוצע במורפמות באמצעות מערכת ה- CHILDES. בשורה העליונה בפלט מופיעה הפקודה שבאמצעותה נעשה החישוב (**MLU -f +t*MOT sample.cha**). תחילה מופיע שם הפקודה (MLU). מיד אחריו מופיע "-f" המציין כי תוצאות הפקודה יופיעו על המסך ולא ישמרו בקובץ. אחר כך מופיע שם הדובר שמבעיו ינותחו (+t*MOT) ושם קובץ התעתיק שינותח (sample.cha). בשורה הבאה מופיע תאריך ביצוע הניתוח והסבר מילולי של פקודת ה-MLU. מתחת לקו ההפרדה מופיע שם קובץ התעתיק המנותח, מתחתיו שם הדובר שמבעיו מנותחים, ואחר כך שלוש שורות שבהן תוצאות הניתוח. ראשית, ספירה של מספר המבעים והמילים, לאחריה היחס ביניהם, ולבסוף סטיית התקן.

תקציר הפקודות שהוזכרו בפרק זה מופיע בטבלה 3, להלן:

סוג הניתוח	פקודה	מהות
חיפוש נתונים	KWAL	רשימות של מבעי הדובר המכילים מילה או צירוף מסוימים
	COMBO	רשימות של מבעים בעלי קידוד זהה
	GEMFREQ	רשימות של קטעי תעתיק מיוחדים
שכיחויות	FREQ	רשימות של מילים ושכיחותן בתעתיק
	COOCCUR	רשימות של צירופים ושכיחותם בתעתיק
מדדים התפתחותיים	MLU	חישוב אורך מבע ממוצע במורפמות או במילים
	MLT	חישוב אורך תור ממוצע
	TTR	חישוב יחס תבנית-תמנית
	VOCD	מדד סטטיסטי לקביעת גיוון לקסיקאלי

טבלה 3: ניתוח נתונים ב-CHILDES

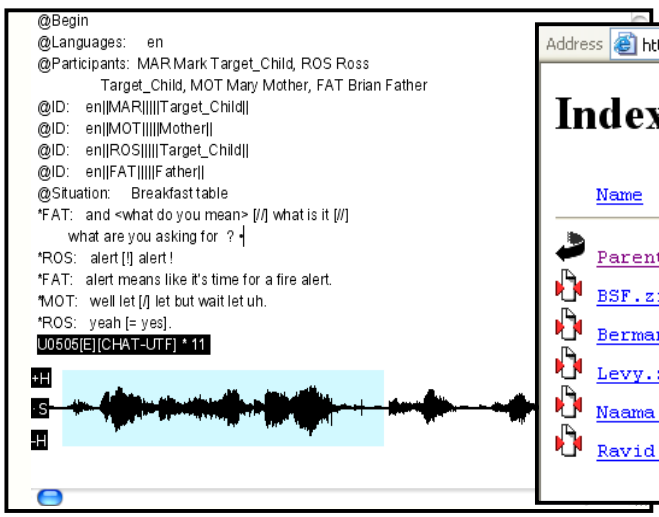
3. קישור לאודיו ולווידאו

מערכת ה-CHILDES מאפשרת לקשר את 'חומרי הגלם' שנאספו לצורך המחקר – קובצי הקול (אודיו) והסרטים (וידאו) – לקובצי התעתיק. לשם כך יש להעביר את עורך הטקסטים למצב של 'תעתוק' אודיו ווידאו. במצב זה ניתן להפעיל סדרה של פקודות שבאמצעותן מקושרת כל שורת תעתיק למקטע שבה היא נשמעת בהקלטה או נראית בסרט. מקטע הקלטה שמתאים לשורת תעתיק מסוימת ניתן לשמוע או לראות באמצעות לחיצה על כפתור מיוחד בשורת התעתיק. היתרון העיקרי של פונקציה זו הוא בכך שהיא מאפשרת לחוקר לבקר את התעתיק על ידי השוואתו להקלטה או לסרט שעליהם הוא מבוסס. תיאור מפורט יותר של אפשרויות הקישור לווידאו ואודיו ניתן למצוא באתר ה-CHILDES בקישורים האלה: אודיו http://talkbank.org/da, וידאו http://www.talkbank.org/dv.

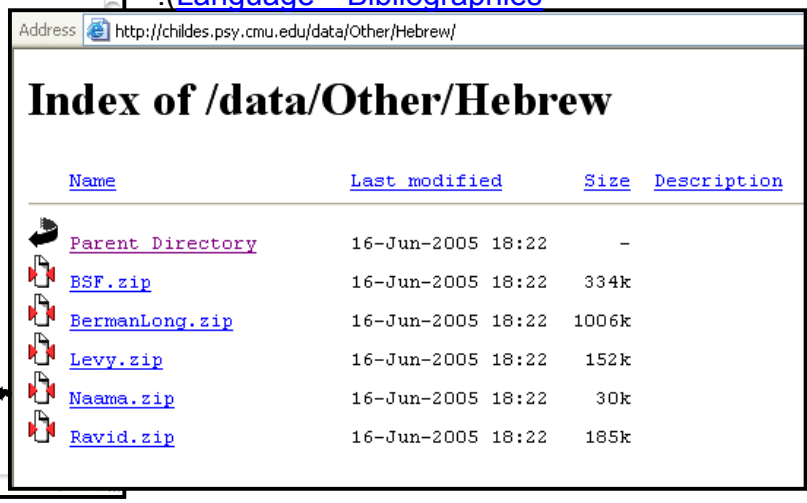
4. עוד מדורי תוכן באתר ה-CHILDES

אתר ה-CHILDES כולל מדורי תוכן מגוונים שהעיקריים שבהם ייסקרו להלן. אחד המדורים החשובים באתר הוא **מדור מסדי הנתונים** (Database). מדור זה כולל תיקיות שבהן קורפוסים שנתרמו למערכת על ידי חוקרים בתחום רכישת השפה על היבטיה השונים. התעתיקים בקורפוסים הללו בנויים בפורמט ה-CHILDES והם זמינים וניתנים להורדה מאתר המערכת ללא תשלום. לכל קורפוס נלווה קובץ תיעוד, המתאר את מספר הקבצים בו וגודלם, את התרחישים שבהם הוא נאסף ופרטים רלוונטיים על הילדים שהוקלטו. כל התעתיקים ניתנים לקידוד וניתוח באמצעות הכלים הייעודיים של המערכת. תרשים 10 מציג רשימה חלקית של הקורפוסים בתחום רכישת העברית. קורפוסים אלה כוללים נתוני אורך וחתך של ילדים דוברי-עברית מן הגיל הרך ועד גיל בית-הספר שהתפתחותם הלשונית תקינה, נתוני חתך של ילדים בעלי לקויות וטקסטים נרטיביים.

עוד מדור בעל חשיבות בעבור חוקר השפה הוא ***מדור הביבליוגרפיה*** (Child Language Bibliographies).



תרשים 9: קישור התעתיק לקובץ קול



תרשים 10: רשימה חלקית של קורפוסים בתחום רכישת העברית

מדור זה כולל רשימה ביבליוגרפית ענפה של כעשרים ושישה אלף מקורות בתחום רכישת השפה. חלק מן המאמרים ברשימה משתמשים במערכת ה-CHILDES ואחרים לא, אולם רובם ככולם עוסקים בהיבטים שונים של רכישת שפה. ניתן לעבור על הרשימה הביבליוגרפית לאחר התקנת

תכנה ייעודית אשר קישור אליה מופיע באתר.

באתר המערכת קיים ***מדור קישורים*** (Other Child Language sites). מדור זה כולל הפניות לאתרים רלוונטיים בתחומים שונים של רכישת שפה וכן לאתרים של ארגונים ואגודות העוסקות בתחום זה ובתחומים קרובים. לצד כל קישור במדור מופיע תיאור תמציתי של האתר או הארגון שאליו הוא מפנה.

לבסוף, בדף הבית של מערכת ה-CHILDES מופיע קישור ל***רשימות הדיון*** של המערכת (Mailing Lists). רשימות אלה משמשות, מצד אחד, במה לשיתוף מידע וסיוע בנושאים טכניים הקשורים לתכנה וביצועיה, ומצד אחר, במה להתדיינות בסוגיות תאורטיות הקשורות ברכישת השפה. באמצעותן ניתן להתעדכן באופן שוטף בשינויים ובעדכונים במערכת ה-CHILDES, ולקבל מידע על כנסים או ספרים חדשים היוצאים לאור בתחומי העניין של קהיליית החוקרים.

5. סיכום

מערכת ה-CHILDES היא מערכת ייעודית לתעתוק, קידוד וניתוח נתונים לשוניים. המערכת, כפי שתוארה כאן, נועדה לאפשר שיתוף בין חוקרים, ומשום כך תעתוק וקידוד הנתונים באמצעותה נעשה על פי סטנדרטים קשיחים. היא ניתנת להורדה ללא תשלום מאתר ה-CHILDES, היא בעלת ממשק ידידותי למשתמש ויכולות נרחבות, וניתן להשתמש במשאביה להוראה ולמחקר. המערכת כוללת כלים מגוונים לעיבוד ***נתונים משפות שונות***, ביניהן שפות שמיות כמו עברית וערבית, ***בתרחישים שונים*** (ניסוי, אינטראקציה ספונטנית), ***מאוכלוסיות שונות*** (מבוגרים, ילדים, דוברים ששפתם תקינה, לקויי-שפה, טעוני טיפוח) ***ואופנויות שונות*** (שפה דבורה או כתובה). המערכת מאפשרת קישור התעתיקים לקובצי אודיו ווידאו. באתר הבית של המערכת ניתן למצוא כמה מדריכים מקוונים (CHAT Transcription system, CLAN Programs, Database Manuals), וכן קישורים למערך התמיכה ופורום הדיון. המערכת דינמית - נוספים לה רכיבים חדשים לעתים קרובות, והיא מתעדכנת תדיר בהתאם לצורכי המשתמשים. גדולתה של המערכת בכך שהיא הופכת חומר מחקרי רב ומגוון בתחום רכישת השפה לנגיש וזמין לקהיליית החוקרים בתחום, ובכך מעודדת שיתוף פעולה בין חוקרים ותורמת לשיפור איכות המחקר הבלשני.

מקורות

Berman, R. A. (1989). Word order project coding (appendix). Unpublished MS, Tel Aviv University.

Berman, R. A. & Weissenborn, J. (1991). **Acquisition of word order: A cross-linguistic study.** Final Report submitted to the German-Israeli Foundation for Scientific Research and Development (G.I.F), Bonn and Jerusalem.

MacWhinney, B. (2000). **The CHILDES project: Tools for analyzing talk**. *Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Uziel-Karl, S. (2001). **A Multidimensional Perspective on the acquisition of Verb Argument Structure**. Unpublished Doctoral Dissertation, Tel Aviv University.