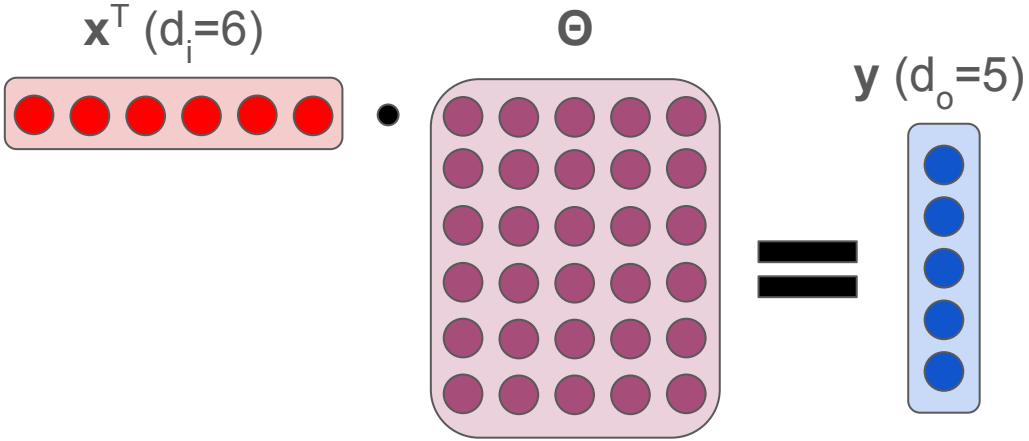
עיבוד שפה טבעית ש9:
תחביר
(SLP 17-18, E 9-11.1)

פתרון הבוחן

השלמה - כפל מטריצות כפעולה ברשת נוירונים

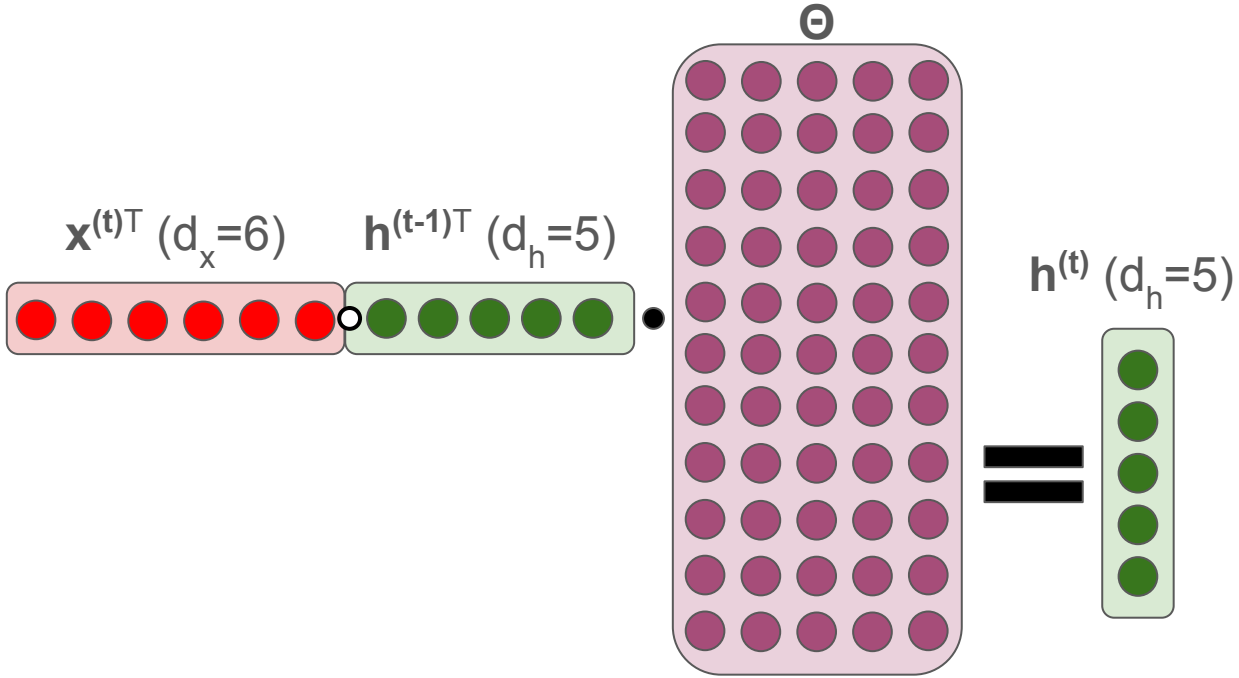
- ברשת FFN רגילה: שכבה מקבלת וקטור **x** ממימד d_i ומוציאה וקטור **y** ממימד d_o
- ולכן **Θ** הוא מטריצה: $\mathbf{y} = \mathbf{x}^T\Theta$)



השלמה - כפל מטריצות כפעולה ברשת נוירונים

- ברשת נשנית: שכבה מקבלת וקטור **x** ממימד d_x
  ווקטור חבוי h ממימד d_h
  ומוציאה וקטור **h** ממימד d_h

- **Θ** נשארת מטריצה:
  $\mathbf{h}^{(t)} = \Theta((\mathbf{x} \circ \mathbf{h}^{(t-1)})^T)$



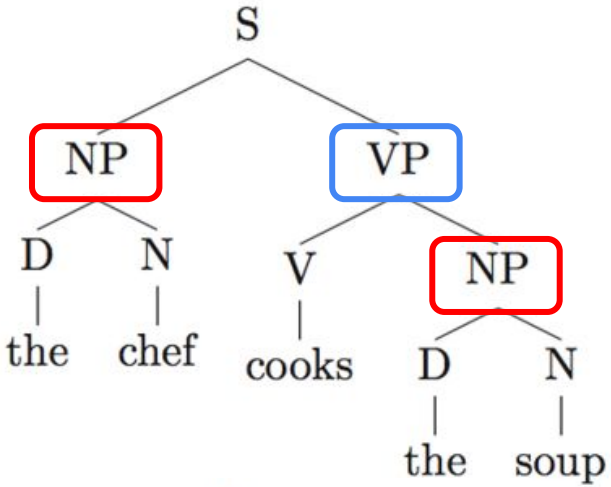
יחידת יסוד תחבירית: צירוף (phrase)

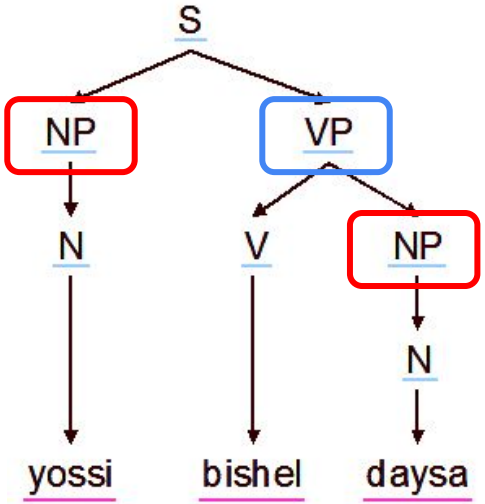
- הילדה אכלה את התפוח
- [הילדה] [אכלה את התפוח]
- [הילדה] [אכלה את [התפוח]]
- [מרב] אכלה את התפוח — צירוף שמני (NP - noun phrase)
- [היא] אכלה את התפוח
- הילדה [נסעה] — צירוף פעלי (VP - verb phrase)
- הילדה אכלה את [זה] — ?

לכל ביטוי יש **ראש תחבירי** לפי הסוג שלו ולפי חלק הדיבר המתאים

תחביר ומבנים

- בדקדוק מרכיבי (**constituency grammar**) ניתן לחלק משפטים ל**צירופים** (phrases) סמוכים ומקוננים
- הצירופים נקראים על-שם ה"ראש", שהוא המרכיב המגדיר אותם לשונית
- אתגר במציאת מבנה המשפט (לאנשים ולאלגוריתמים) - דו-משמעות
  - לקסיקלית - מילה יכולה לבטא כמה משמעויות, או כמה חלקי דיבר (חולצה מטיילת בואדי)
  - תחבירית (מבנית) - אותן מילים, אותם חלקי דיבר, פרשנות מבנה שונה (⇐)





דוגמאות לדו-משמעות מבנית

- הממשלה מבטלת את התו הירוק בבריכות, במסעדות ובאירועים זועמים.
- כסא אופניים לילד שמתחבר לסבל
- מכתב שכתבה מורה לתיאטרון

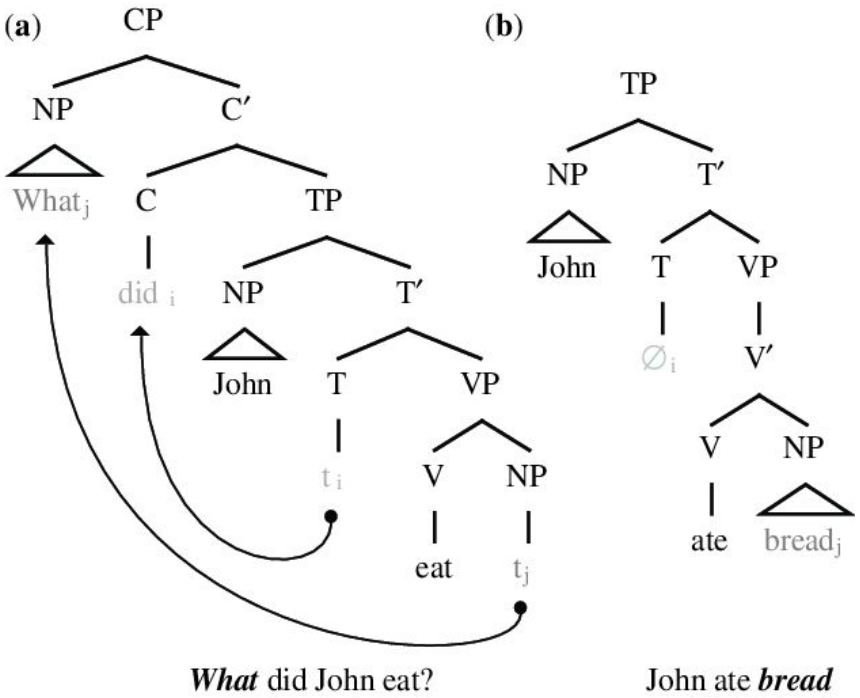
תחביר כהסבר: קורפרנס (coreference), אנאפורה (anaphora)

- (האינדקסים i,j מציינים שמדובר באותו אדם)

- שפרה$_i$ מדברת אליה$_j$.
- * שפרה$_i$ מדברת אליה$_i$.
- שפרה$_i$ מדברת אל עצמה$_i$.
- * היא$_i$ מדברת לשפרה$_i$.
- שפרה$_i$ אוהבת שהיא$_j$ לפעמים שותקת.
- שפרה$_i$ אוהבת שהיא$_i$ לפעמים שותקת.
- שפרה$_i$ ביקשה מאגאתה$_j$ לעזור לה$_i$.
- * שפרה$_i$ ביקשה מאגאתה$_j$ לעזור לעצמה$_i$.
- שפרה$_i$ ביקשה מאגאתה$_j$ לעזור לעצמה$_j$.

- (מעוניינים בהסברים בלשניים? גגלו Binding theory)

Movement "תנועה"



What did John eat?

John ate ***bread***

תחביר כהסבר (II): מגבלות על תנועה

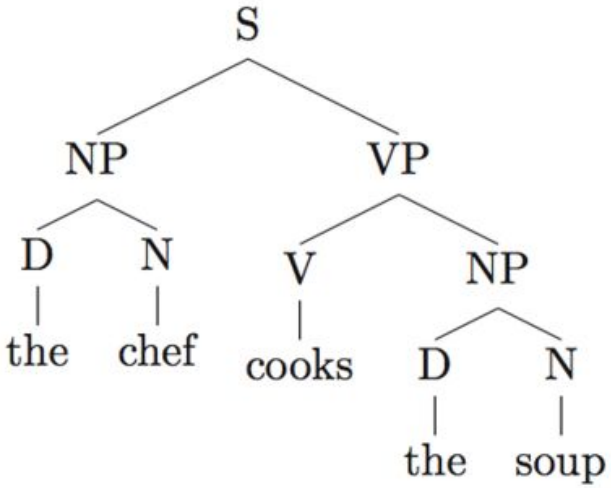
- יוסי בישל דיסה.
  - מי בישל דייסה?
  - מה יוסי בישל?
  - את מי יוסי פגש?
- רינה טענה שיוסי בישל דיסה.
  - מי רינה טענה שבישל דיסה?
  - מה רינה טענה שיוסי בישל?
- רינה קבעה את הקביעה שיוסי בישל דיסה.
  - * מה רינה קבעה את הקביעה שיוסי בישל?
- Rina bought Yossi’s book
  - Whose book did Rina buy?
  - * Whose did Rina buy book?
- (מעוניינותים בהסברים בלשניים? גגלו Ross's Islands)

דקדוק חסר-הקשר (CFG - Context Free Grammar)

- אמצעי "מעלה-מטה" (top-down) להתחיל מצומת השורש ולאט לאט **לגזור** עוד צמתים עד שמגיעים למשפט
  - גנרטיבי **מאוד**
- כלל המשפטים החוקיים בשפה נקבעים ע"י **כללי הגזירה**, שיכולים להיות רקורסיביים
- פורמלית:

צירופים ותגים	N a set of **non-terminal symbols** (or **variables**)
מילים	Σ a set of **terminal symbols** (disjoint from N)
כללי גזירה	R a set of **rules** or productions, each of the form $A \rightarrow \beta$, where A is a non-terminal, β is a string of symbols from the infinite set of strings $(\Sigma \cup N)^*$
שורש	S a designated **start symbol** and a member of N

כללי גזירה



- S → NP VP
- NP → D N
- VP → V NP
- D → the
- N → chef || soup
- V → cooks

- (בדקדוק הזה יש רקורסיה?)

- צורת חומסקי נורמלית (**CNF**): כל כלל כולל מימין או (1) שני נונטרמינלים (nonterminals), או (2) מילה אחת.
  - (האם הדקדוק לעיל הוא CNF?)

עוצמת CFG [דילגנו בשיעור]

- יותר כללי גזירה - יותר עצים אפשריים
- אבל, כמה כללי גזירה צריך?
- כל צורת ביטוי אפשרי בשפה דורשת עוד כלל

Grammar	Lexicon
$S \rightarrow NP\ VP$	$Det \rightarrow that \mid this \mid the \mid a$
$S \rightarrow Aux\ NP\ VP$	$Noun \rightarrow book \mid flight \mid meal \mid money$
$S \rightarrow VP$	$Verb \rightarrow book \mid include \mid prefer$
$NP \rightarrow Pronoun$	$Pronoun \rightarrow I \mid she \mid me$
$NP \rightarrow Proper\text{-}Noun$	$Proper\text{-}Noun \rightarrow Houston \mid NWA$
$NP \rightarrow Det\ Nominal$	$Aux \rightarrow does$
$Nominal \rightarrow Noun$	$Preposition \rightarrow from \mid to \mid on \mid near \mid through$
$Nominal \rightarrow Nominal\ Noun$	
$Nominal \rightarrow Nominal\ PP$	
$VP \rightarrow Verb$	
$VP \rightarrow Verb\ NP$	
$VP \rightarrow Verb\ NP\ PP$	
$VP \rightarrow Verb\ PP$	
$VP \rightarrow VP\ PP$	
$PP \rightarrow Preposition\ NP$	

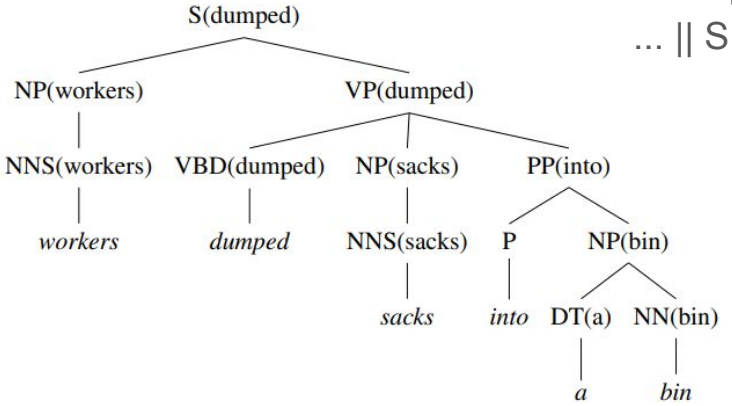
Figure 13.1 The $\mathscr{L}_1$ miniature English grammar and lexicon.

http://ironcreek.net/syntaxtree/

לקסיקליזציה (אייזנסטיין 10.5.2)

- דקדוק מועשר בידע מילוני

- מסגרת סבקטגוריזציה subcategorization frame - "מה המילה מקבלת כמשלים"
  - *עמדה*: נושא NP
  - *כתבה*: נושא NP (+ מושא ישיר NP)
  - *נתנה*: נושא NP + מושא ישיר NP (+ מושא עקיף PP-ל)
  - *רצתה*: נושא NP + מושא ישיר NP || נושא NP + מושא פסוקי משועבד SBAR
  - *אמרה*: נושא NP + מושא עקיף NP-ל + מושא פסוקי לא נטוי S || ...



- תת-אפיון של הצירופים
  - NP-SBJ - צירוף בעמדת נושא, NP-OBJ בעמדת מושא, וכו'

- סימון ראש הצירוף על-גבי העץ

אוניברסליות?

- כבר באנגלית, הרבה תופעות צריכות הסברים בנוסח "תנועה" או "אלמנטים ריקים"
  - What did the man want?
  - I brought the notebook to class and the laptop to the office
- בשפות עם סדר מילים יותר חופשי, קשה לקבץ הכל בצירופים או לתרץ את המבנים
- בשפות שבהן כללי התאם חשובים, לפעמים קשה להסיק עליהם מתוך המבנים

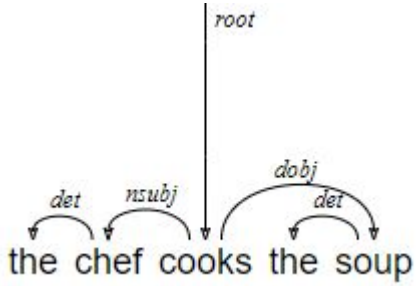
מבנה תלויות (Dependency structure) (אייזנסטיין 11.1)

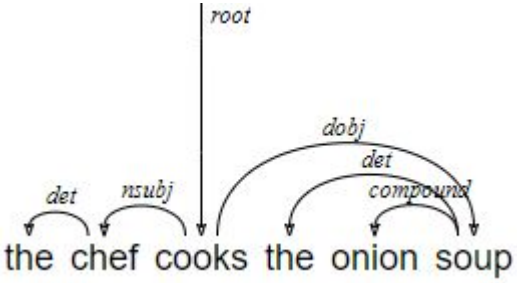
- נתאר ישירות קשר בין מילים במשפט ונבנה היררכיה שבה רק המילים הן צמתים

the chef cooks the soup

מבנה תלויות (Dependency structure) (אייזנסטיין 11.1)

- נתאר ישירות קשר בין מילים במשפט ונבנה היררכיה שבה רק המילים הן צמתים
- בפועל, נדבר על **עץ תלויות** (**dependency tree**)
- קשת מוגדרת ע"י:
  - המקור שלה, או ראשה **head**
  - היעד שלה, או תלויה **dependent**
  - הסוג שלה **type / label / relation**





סוגי קשתות בעץ תלויות

- קשתות בין נשוא (predicate) לארגומנט ליבה (core argument)
  - נושא שמני nsubj
  - נושא פסוקי csubj
  - מושא ישיר dobj
  - מושא עקיף iobj
  - פסוקית משלימה ccomp, xcomp
- קשתות בין נשוא למרכיבי עזר (non-core dependent)
  - נספחים שמניים obl
  - פסוקית אופן advcl
  - פועל עזר aux, cop

סכימות התיוג מתעדכנות עם השנים

כשלא מוגבלים לעצים, כאן ייתכנו קשתות נכנסות כפולות

https://universaldependencies.org/u/dep/index.html

סוגי קשתות בעץ תלויות

- קשתות בין נשוא (predicate) לארגומנט ליבה (core argument)
- קשתות בין נשוא למרכיבי עזר (non-core dependent)
- קשתות בתוך צירופים שמניים
  - תווית יידוע determiner
  - סמיכות nmod
  - שם תואר amod
  - מילת יחס או יחסה case
- אחרות
  - חיבור conj, cc, list
  - "מיוחדים" goeswith

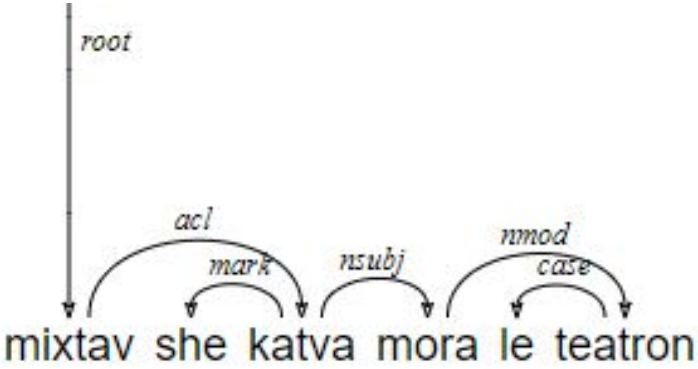
סכימות התיוג מתעדכנות עם השנים

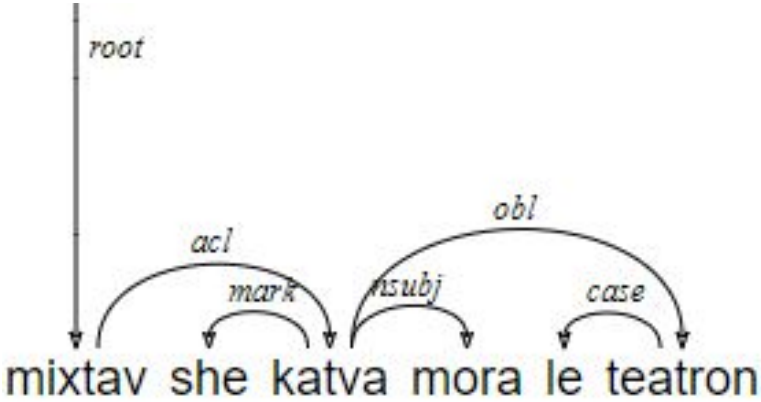
כשלא מוגבלים לעצים, כאן ייתכנו קשתות נכנסות כפולות

https://universaldependencies.org/u/dep/index.html

דו-משמעות בעץ תלויות

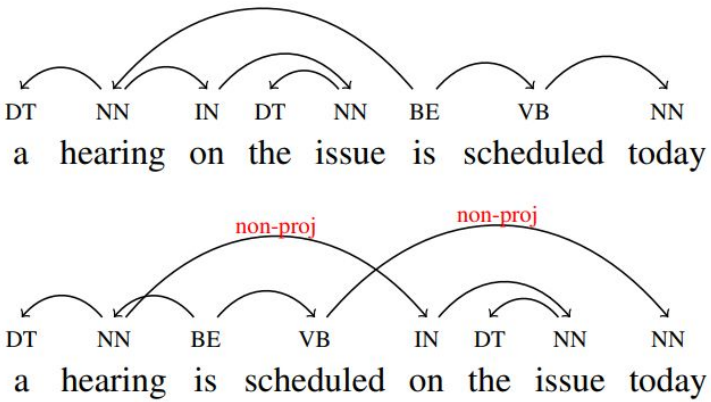
- מכתב שכתבה מורה לתיאטרון





היטליות (Projectivity)

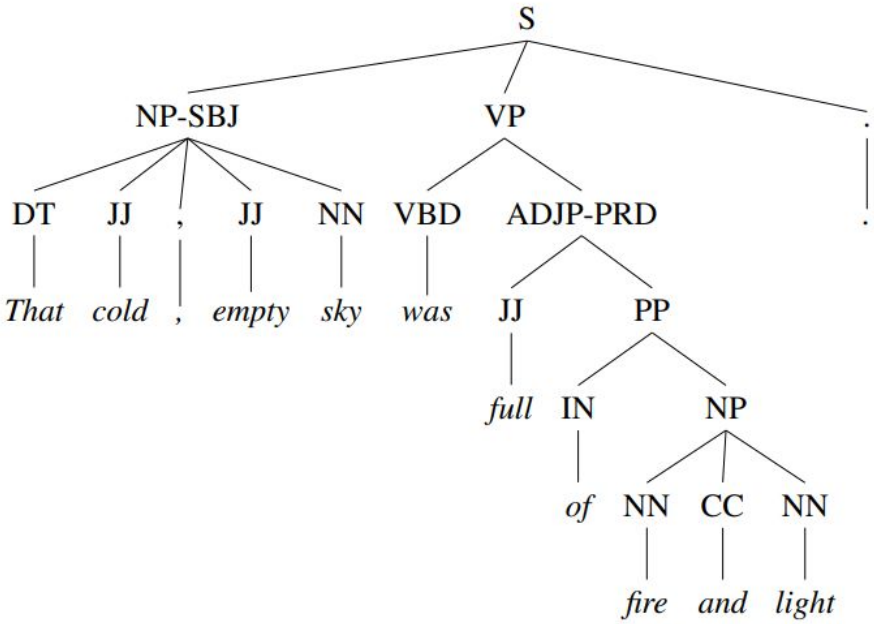
- התכונה לפיה קשתות העץ **לא חוצות זו את זו**
  - התייחסות יחידה ל*סדר המילים* בדקדוק תלויות
  - האם יש התייחסות כזו בדקדוק מרכיבי?
- האם יש בשפה טבעית משפטים שהעץ שלהם אינו היטלי?
- הגדרה פורמלית: קשת מ-h ל-d היא היטלית ⇔ כל המילים שבין h ו-d ברצף המילים הן צאצאיות של h בגרף; עץ הוא היטלי ⇔ כל הקשתות בו היטליות



משאבים (מרכיבי)

- אוסף העצים של פן (Penn Treebank, PTB)

```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```



משאבים (תלויות)

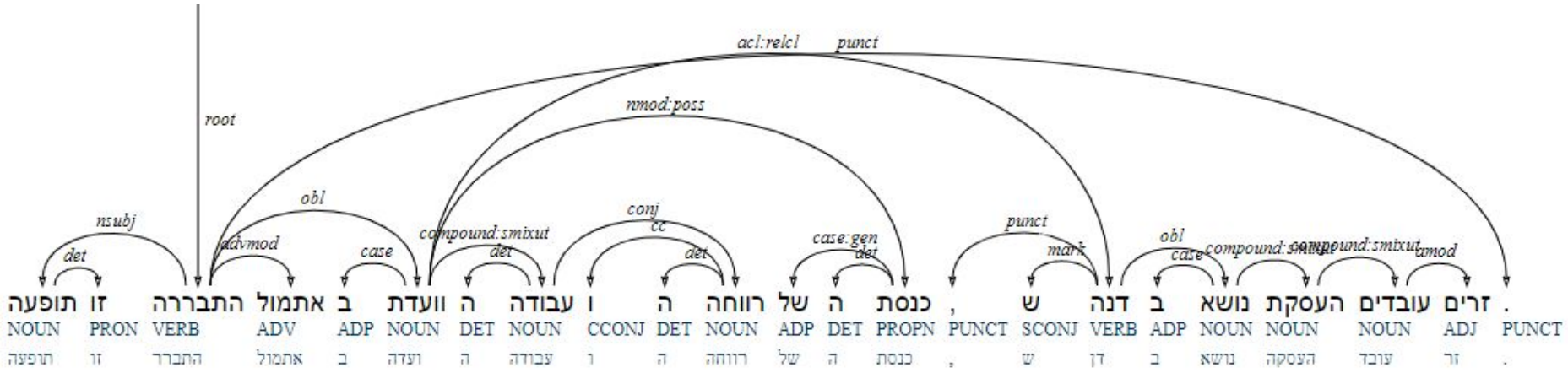
- תלויות אוניברסליות (UD) Universal Dependencies

Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	<1K		Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	2	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Apurina	1	<1K		Arawakan
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	2	55K		IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Beja	1	1K		Afro-Asiatic, Cushitic
Belarusian	1	305K	W	IE, Slavic
Bengali	2	<1K	W	IE, Indic
Bhojpuri	2	6K		IE, Indic
Breton	1	10K	W	IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic
Cantonese	1	13K		Sino-Tibetan
Catalan	1	553K		IE, Romance
Chinese	5	285K	W	Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Chinese	1	283K		Sino-Tibetan
Coptic	1	52K		Afro-Asiatic, Egyptian
Croatian	1	199K	W	IE, Slavic
Czech	6	3,428K	W	IE, Slavic
Danish	2	100K		IE, Germanic
Dutch	2	306K	W	IE, Germanic
English	11	1,943K	W	IE, Germanic
Erzya	1	17K		Uralic, Mordvin
Estonian	2	511K		Uralic, Finnic
Faroese	2	50K	W	IE, Germanic
Finnish	4	397K	W	Uralic, Finnic
French	9	1,195K	W	IE, Romance
Frisian	1	51K	W	IE, Germanic
Frisian Dutch	1	3K		Code switching
Galician	2	164K		IE, Romance
German	4	3,810K	W	IE, Germanic
Gothic	1	55K		IE, Germanic
Greek	1	63K	W	IE, Greek
Guajajara	1	2K		Tupian, Tupi-Guarani
Hebrew	1	161K		Afro-Asiatic, Semitic
Hindi	2	375K	W	IE, Indic
Hindi English	1	26K		Code switching
Hittite	1	1K		IE, Anatolian
Hungarian	1	42K		Uralic, Ugric
Icelandic	3	1,162K	W	IE, Germanic
Indonesian	3	169K	W	Austronesian, Malayo-Sumbawan
Irish	2	131K		IE, Celtic
Italian	7	818K	W	IE, Romance
Japanese	8	2,849K	W	Japanese
Javanese	1	2K		Austronesian, Javanese
Kaapor	1	<1K		Tupian, Tupi-Guarani
Kabyle	1	47K		Afro-Asiatic, Berber
Kangri	1	2K		IE, Indic
Karelian	1	3K		Uralic, Finnic
Karo	1	2K		Tupian, Purubora-Ramarama
Kazakh	1	10K	W	Turkic, Northwestern
Khunsari	1	<1K		IE, Iranian
Kiche	1	10K	W	Mayan
Komi Permyak	1	<1K		Uralic, Permic
Komi Zyrian	2	10K		Uralic, Permic
Korean	5	446K	W	Korean
Kurmanji	1	10K	W	IE, Iranian
Latin	5	977K		IE, Latin
Latvian	1	265K		IE, Baltic
Laz	1	2K		Kartvelian
Ligurian	1	6K	W	IE, Romance
Lithuanian	2	75K		IE, Baltic
Livvi	1	1K		Uralic, Finnic
Low Saxon	1	2K		IE, Germanic
Magahi	2	7K	W	IE, Indic
Makurap	1	<1K		Tupian, Tupari
Maltese	1	44K	W	Afro-Asiatic, Semitic
Manx	1	20K	W	IE, Celtic
Marathi	1	3K	W	IE, Indic
Mbya Guarani	2	13K		Tupian, Tupi-Guarani
Middle Irish	2	<1K		IE, Celtic
Moksha	1	3K		Uralic, Mordvin
Munduruku	1	<1K		Tupian, Munduruku
Naija	1	140K		Creole
Nayini	1	<1K		IE, Iranian
Neapolitan	1	<1K		IE, Romance
North Sami	1	26K		Uralic, Sami
Norwegian	3	666K		IE, Germanic
Odia	1	<1K		IE, Indic
Old Church Slavonic	1	57K		IE, Slavic
Old East Slavic	3	184K		IE, Slavic
Old French	1	199K		IE, Romance
Old Irish	2	21K		IE, Celtic
Old Japanese	1	<1K		Japanese
Old Turkish	1	<1K		Turkic, Northeastern
Persian	2	654K		IE, Iranian
Polish	3	499K		IE, Slavic
Portuguese	4	570K	W	IE, Romance
Prakrit	1	<1K		IE, Indic
Romanian	4	937K	W	IE, Romance
Russian	4	1,832K	W	IE, Slavic
Sanskrit	2	28K		IE, Indic
Scottish Gaelic	1	84K		IE, Celtic
Serbian	1	97K		IE, Slavic
Sindhi	1	6K		IE, Indic
Skolt Sami	1	2K		Uralic, Sami
Slovak	1	106K		IE, Slavic
Slovenian	2	170K		IE, Slavic
Soi	1	<1K		IE, Iranian
South Levantine Arabic	1	<1K		Afro-Asiatic, Semitic
Spanish	3	1,022K	W	IE, Romance
Swedish	3	206K	W	IE, Germanic
Swedish Sign Language	1	1K		Sign Language
Swiss German	1	1K	W	IE, Germanic
Tagalog	2	1K		Austronesian, Central Philippine
Tamil	2	12K		Dravidian, Southern
Tatar	1	1K		Turkic, Northwestern
Telugu	1	6K		Dravidian, South Central
Thai	1	22K	W	Tai-Kadai
Tupinamba	1	1K		Tupian, Tupi-Guarani
Turkish	9	733K	W	Turkic, Southwestern
Turkish German	1	37K		Code switching
Ukrainian	1	122K	W	IE, Slavic
Upper Sorbian	1	11K	W	IE, Slavic
Urdu	1	138K		IE, Indic
Uyghur	1	40K		Turkic, Southeastern
Vietnamese	1	43K		Austro-Asiatic, Viet-Muong
Warlpiri	1	<1K		Pama-Nyungan
Welsh	1	41K	W	IE, Celtic
Western Armenian	1	92K		IE, Armenian
Wolof	1	44K	W	Niger-Congo, Northern Atlantic
Xibe	1	15K		Tungusic
Yakut	1	<1K		Turkic, Northeastern
Yoruba	1	8K	W	Niger-Congo, Defoid
Yupik	1	2K		Eskimo-Aleut

פורמט CoNLL-U

```
# sent_id = 2
# text = תופעה זו התבררה אתמול בוועדת העבודה והרווחה של הכנסת, שדנה בנושא העסקת עובדים זרים.
1	תופעה	תופעה	NOUN	NOUN	Gender=Fem|Number=Sing	3	nsubj	_	_
2	זו	זו	PRON	PRON	Gender=Fem|Number=Sing|Person=3|PronType=Dem	1	det	_	_
3	התבררה	התברר	VERB	VERB	Gender=Fem|HebBinyan=HITPAEL|Number=Sing|Person=3|Tense=Past	0	root	_	_
4	אתמול	אתמול	ADV	ADV	_	3	advmod	_	_
5-6	בוועדת	_	_	_	_	_	_	_	_
5	ב	ב	ADP	ADP	_	6	case	_	_
6	וועדת	ועדה	NOUN	NOUN	Definite=Cons|Gender=Fem|Number=Sing	3	obl	_	_
7-8	העבודה	_	_	_	_	_	_	_	_
7	ה	ה	DET	DET	Definite=Def|PronType=Art	8	det	_	_
8	עבודה	עבודה	NOUN	NOUN	Gender=Fem|Number=Sing	6	compound:smixut	_	_
9-11	והרווחה	_	_	_	_	_	_	_	_
9	ו	ו	CCONJ	CCONJ	_	11	cc	_	_
10	ה	ה	DET	DET	Definite=Def|PronType=Art	11	det	_	_
11	רווחה	רווחה	NOUN	NOUN	Gender=Fem|Number=Sing	8	conj	_	_
12	של	של	ADP	ADP	Case=Gen	14	case:gen	_	_
13-14	הכנסת	_	_	_	_	_	_	_	SpaceAfter=No
13	ה	ה	DET	DET	Definite=Def|PronType=Art	14	det	_	_
14	כנסת	כנסת	PROPN	PROPN	_	6	nmod:poss	_	_
15	,	,	PUNCT	PUNCT	_	17	punct	_	_
16-17	שדנה	_	_	_	_	_	_	_	_
16	ש	ש	SCONJ	SCONJ	_	17	mark	_	_
17	דנה	דן	VERB	VERB	Gender=Fem|HebBinyan=PAAL|Number=Sing|Person=3|Tense=Past|Voice=Act	6	acl:relcl	_	_
18-19	בנושא	_	_	_	_	_	_	_	_
18	ב	ב	ADP	ADP	_	19	case	_	_
19	נושא	נושא	NOUN	NOUN	Definite=Cons|Gender=Masc|Number=Sing	17	obl	_	_
20	העסקת	העסקה	NOUN	NOUN	Definite=Cons|Gender=Fem|Number=Sing	19	compound:smixut	_	_
21	עובדים	עובד	NOUN	NOUN	Gender=Masc|Number=Plur	20	compound:smixut	_	_
22	זרים	זר	ADJ	ADJ	Gender=Masc|Number=Plur	21	amod	_	SpaceAfter=No
23	.	.	PUNCT	PUNCT	_	3	punct	_	_
```

פורמט CoNLL-U

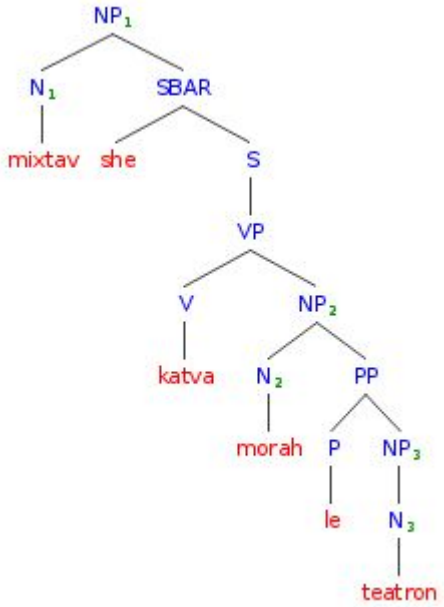


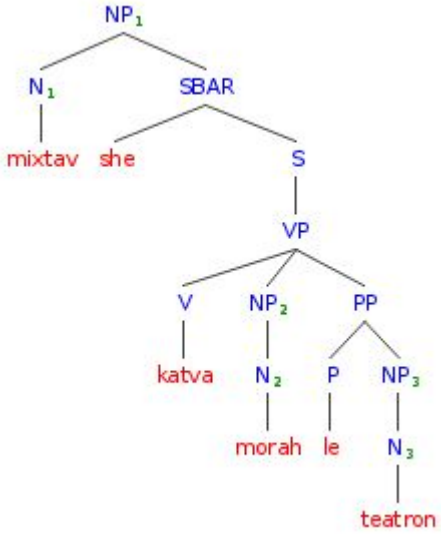
https://arborator.ilpga.fr/q.cgi

מטריקות (מרכיבי) [דילגנו בשיעור]

- פרסבל PARSEVAL: לוקחים את כל המרכיבים בעץ ה"נכון" (gold) ובודקים כמה המערכת זיהתה נכון לגמרי (recall), לוקחים את מרכיבי העץ החזוי ובודקים כמה מהם נכונים (precision), מדווחים F1

- חציית סוגריים (cross-bracketing): יחס המרכיבים **הנכונים** בלי "הפרעה במבנה": סוגר [בלתי סגור בתוך [ ]



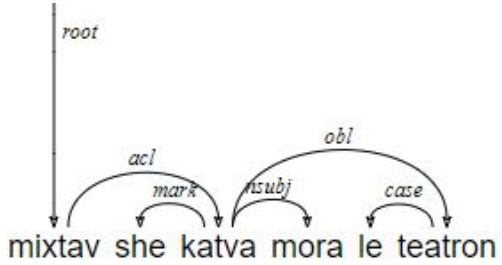


מטריקות (תלויות)

- ציונים פר-מילה (ראשה או תלויה?)
  - עם חשיבות לסוג הקשת LAS - Labeled Attachment Score
  - בלי חשיבות לסוג הקשת UAS - Unlabeled Attachment Score
  - רק סוג הקשת, בלי חשיבות לראשה LS - Label Score
    - איזו מין בעיה זו?

- אינטואיציה: כמה תקבל מערכת על משפט שפורסר נכון חוץ משגיאה בתליית צירוף יחס?

- עוד אפשרות: דיוק פר סוג קשת (לאורך הדאטאסט)
  - מטריקות?



ניתוח תחבירי Syntactic Parsing

- מחר. בינתיים:
- כמה עצי מרכיבים אפשריים יש למשפט בן n מילים?
- כמה עצי CNF אפשריים יש למשפט בן n מילים?
- כמה עצי תלויות אפשריים יש למשפט בן n מילים?
- כמה עצי תלויות היטליים אפשריים יש למשפט בן n מילים?