CLPsych

Hebrew Psychological Lexicons

לקסיקונים פסיכולוגיים בעברית

Natalie Shapira

Dana Atzil-Slonim, Daniel Juravski, Moran Baruch, Adar Paz, Dana Stolowicz-Melman, Tal Alfi-Yogev, Roy Azoulay, Adi Singer, Maayan Revivo, Chen Dahbash, Limor Dayan, Tamar Naim, Lidar Gez, Boaz Yanai, Adva Maman, Adam Nadaf, Elinor Sarfati, Amna Baloum, Tal Naor, Ephraim Mosenkis, Matan Kenigsbuch, Badreya Sarsour, Yarden Elias, Liat Braun, Moria Rubin, Jany Gelfand Morgenshteyn, Noa Bergwerk, Noam Yosef, Sivan Peled, Coral Avigdor, Rahav Obercyger, Rachel Mann, Tomer Alper, Inbal Beka, Ori Shapira, Yoav Goldberg

June 2021

BIU NLP

Psychotherapy Research Lab

DSI Understanding the world through Data Bar-Ilan University
DATA SCIENCE INSTITUTE

Computer Science and Psychology Departments

Supervisors →

Prof. Yoav Goldberg

Dr. Dana Atzil-Slonim

Leading →

Natalie Shapira

Interning Therapists and Psychotherapy →

Adar Paz

Dana Stolowicz-Melman

Tal Alfi-Yogev

Roy Azoulay

Natural Language Processing and Machine Learning →

Ori Shapira

Moran Baruch

Daniel Juravski

Inbal Beka

Psychology Research Practicum → ↓

Adi Singer Ruskin

Maayan Revivo

Chen Dahbash

Limor Dayan

Tamar Naim

Lidar Gez

Boaz Yanai

Adva Maman

Elinor Sarfati

Amna Baloum

Tal Naor

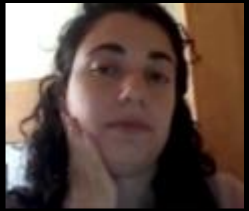
Ephraim Mosenkis

Matan Kenigsbuch

Badreya Sarsour

Yarden Elias

Liat Braun


Moria Rubin

Jany Gelfand Morgenshteyn

Noa Bergwerk

Noam Yosef

Sivan Peled

Coral Avigdor

Rahav Obercyger

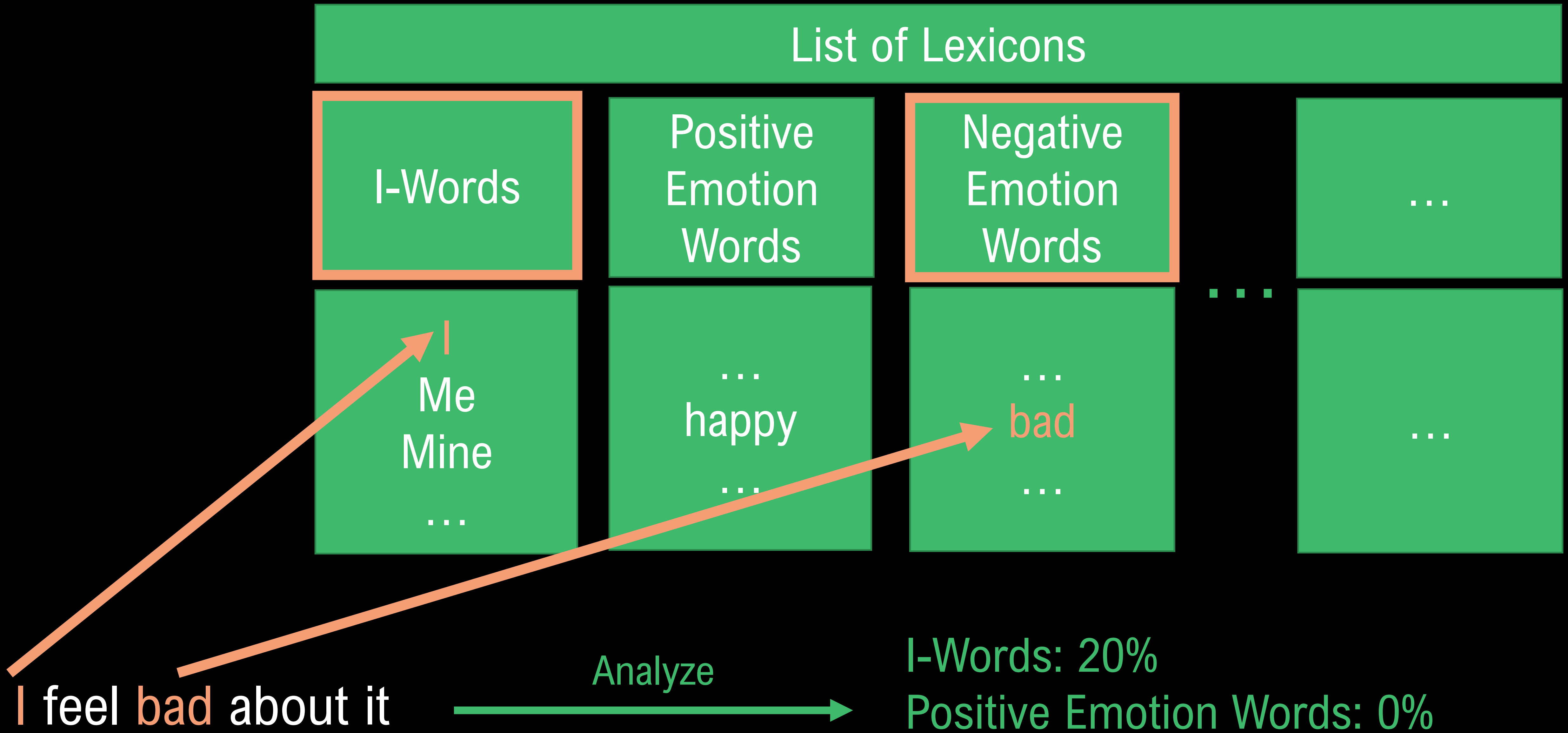
Rachel Mann

Tomer Alper

Adam Nadaf

Outline

- What are lexicons
- Motivation
- Introduce large set of Hebrew lexicons
- Challenges in creating and validating lexicons
- Methodological considerations in lexicon construction process
  - Base Dataset
  - Construction process of each collection and initial results of research studies



List of Lexicons

I-Words

Positive Emotion Words

Negative Emotion Words

…

…

I
Me
Mine
…

…
happy
…

…
bad
…

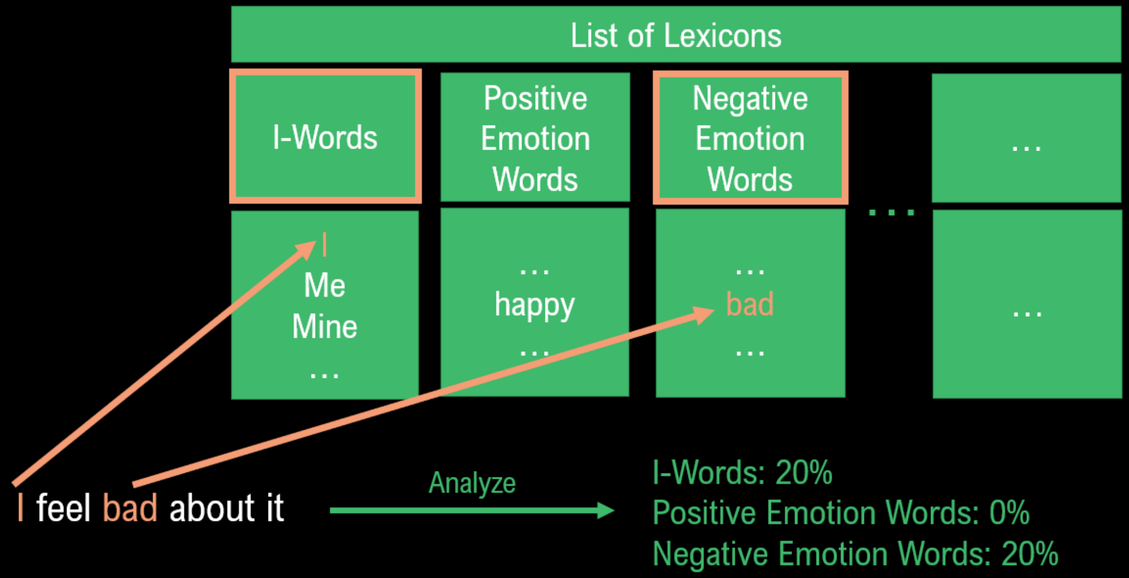
…

I feel bad about it

Analyze

I-Words: 20%
Positive Emotion Words: 0%
Negative Emotion Words: 20%

Why Do We Need Lexicons?



- Scarce data
  - Few samples are available in clinical trials
  - Confidentiality limits sharing of data
- Serve as clinical markers
- Interpretation of results
- Easy to use and improve performance

data-hungry models are not practical in such cases

Collections

Lexicons & Word-Lists

Collection Name (# Lexicons or Lists, # Words)

- Valence (2, 200)
- Emotional Variety (42, 7313)
- Paralinguistics (11, 154)
- Depressive Characteristics (14, 194)
- Well-Being (2, 40)
- Conversation Topics (200, 4000)
- Hebrew LIWC (under construction)
- Extended Emotional Variety (under construction)

Lexicons are freely available at https://github.com/natalieShapira/HebrewPsychologicalLexicons
Hebrew LIWC is for internal use only as LIWC is commercial.

Construction Methods

- Expert Knowledge Based Lexicons
- Data-Driven Lists
  - Supervised
  - Unsupervised
- Expert Knowledge + Automatic Methods
  - Translation
  - Expansion

Validity & Reliability

Coverage

Domain expert verification

Initial research use case

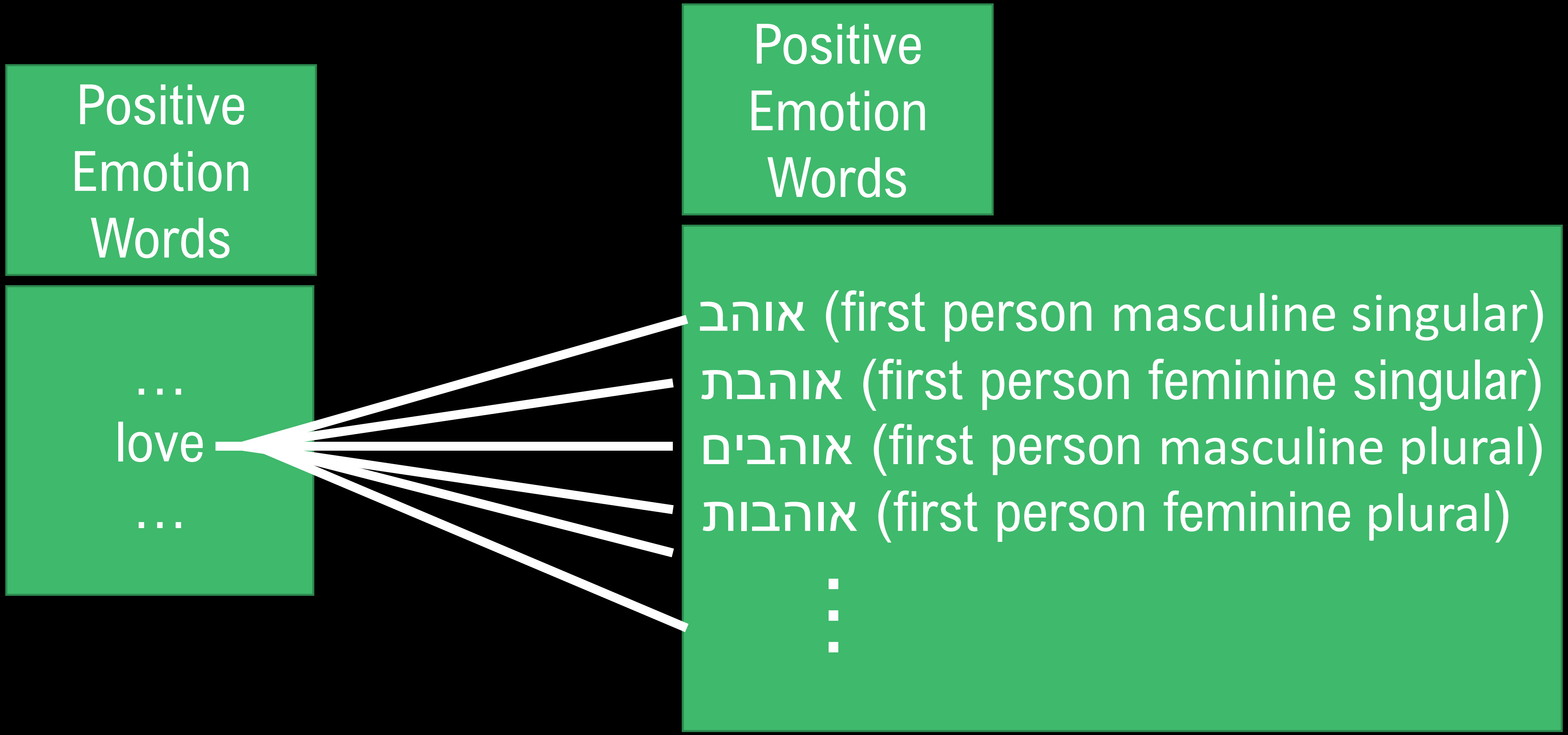
- Outcome Rating Scale (ORS; Miller et al., 2003)
- Profile of Mood States (POMS; McNair, 1992)
- Post-Session Questionnaire (PSQ; Muran et al., 2004)

Challenges with Lexicon Translation

The Challenge - Hebrew

- No LIWC version in Hebrew
- Languages behave differently
- Morphology makes things harder

Morphology
(1) Verb inflections



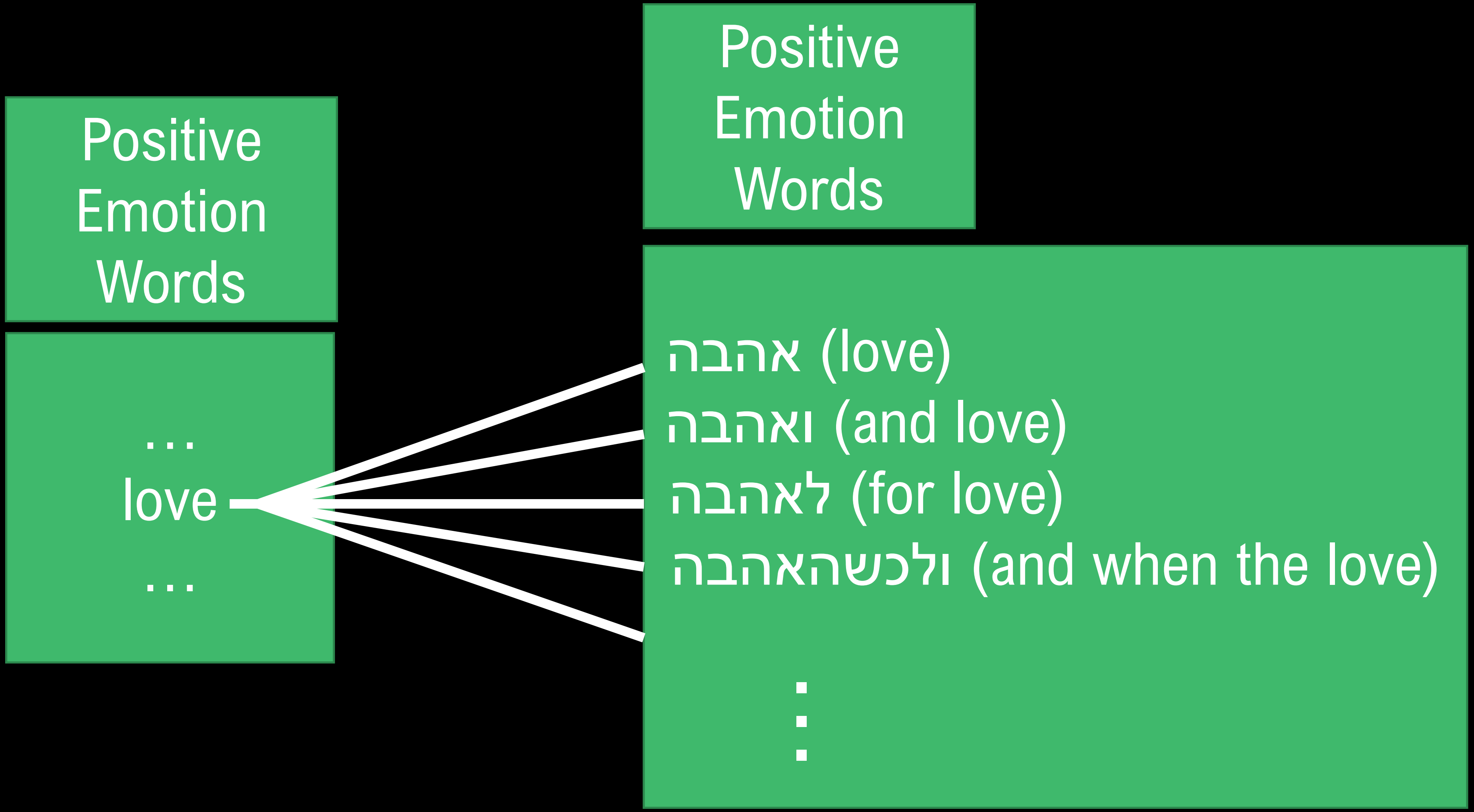
Positive Emotion Words

…
love
…

Positive Emotion Words

אוהב (first person masculine singular)
אוהבת (first person feminine singular)
אוהבים (first person masculine plural)
אוהבות (first person feminine plural)
⋮

Morphology
(2) Clitics / Morphemes



Positive Emotion Words

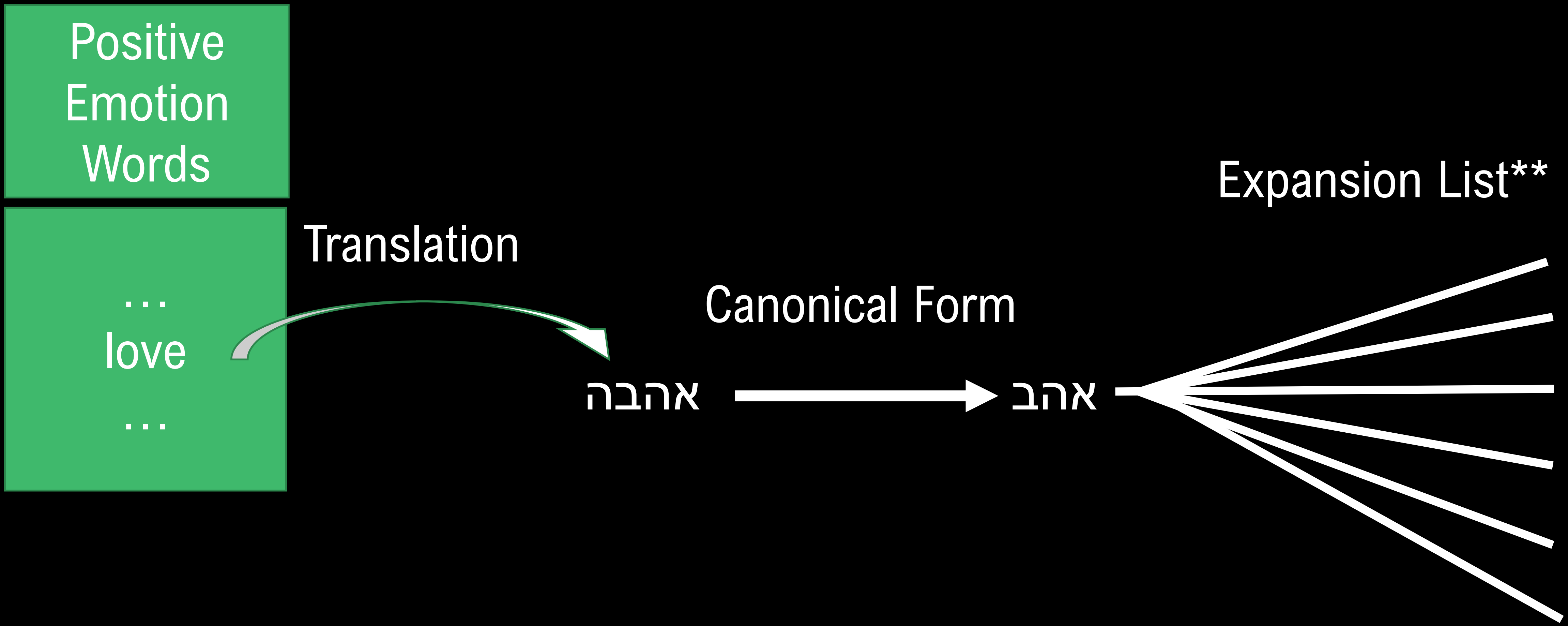
…
love
…

Positive Emotion Words

אהבה (love)
ואהבה (and love)
לאהבה (for love)
ולכשהאהבה (and when the love)

⋮

Solution



Positive Emotion Words

…

love

…

Translation

אהבה

Canonical Form

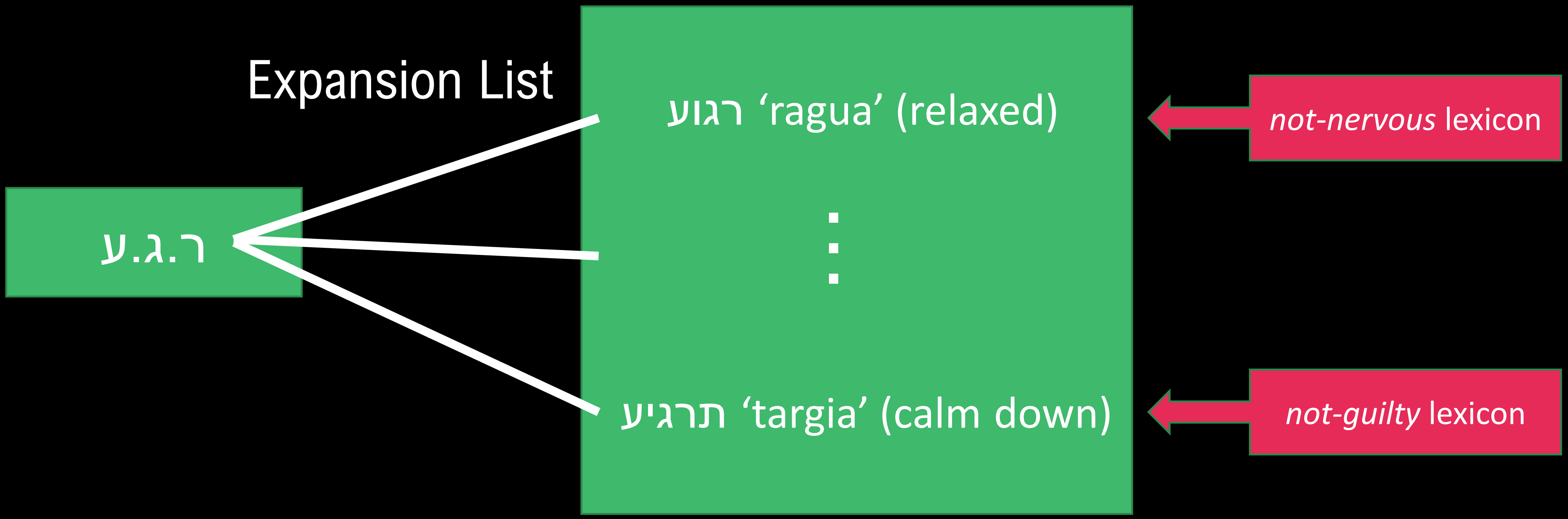
אהב

Expansion List**

** Such as the MILA lexicon (Itai and Wintner, 2008) or the BGU-version of the lexicon, which is bundled with the YAP Hebrew parser (More and Tsarfaty, 2016) as the file bgulex.utf8.hr.

Should all forms of a word should be included in the same lexicon ?

Should all forms of a word should be included in the same lexicon ?
- No.



Expansion List

ר.ג.ע

רגוע ‘ragua’ (relaxed)

⋮

תרגיע ‘targia’ (calm down)

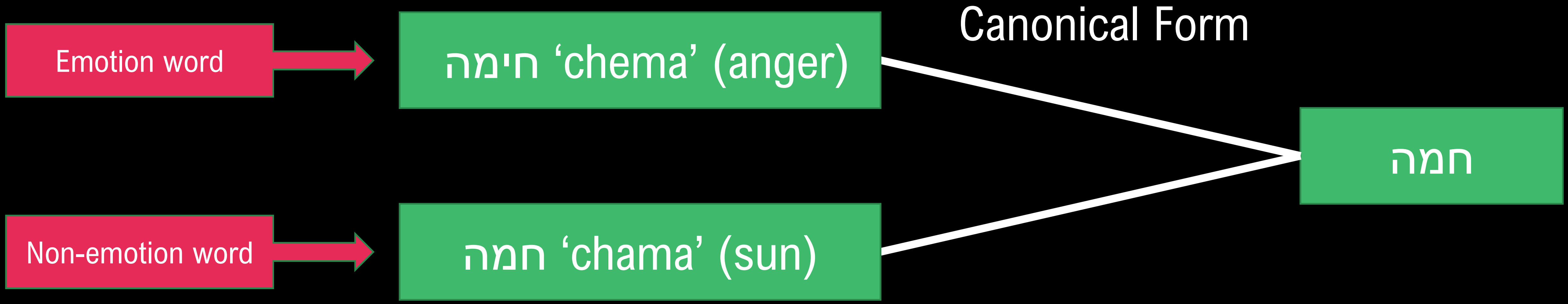
not-nervous lexicon

not-guilty lexicon

Do we keep critical information while converting to canonical form?

Do we keep critical information while converting to canonical form?

- No.



Canonical Form

Emotion word

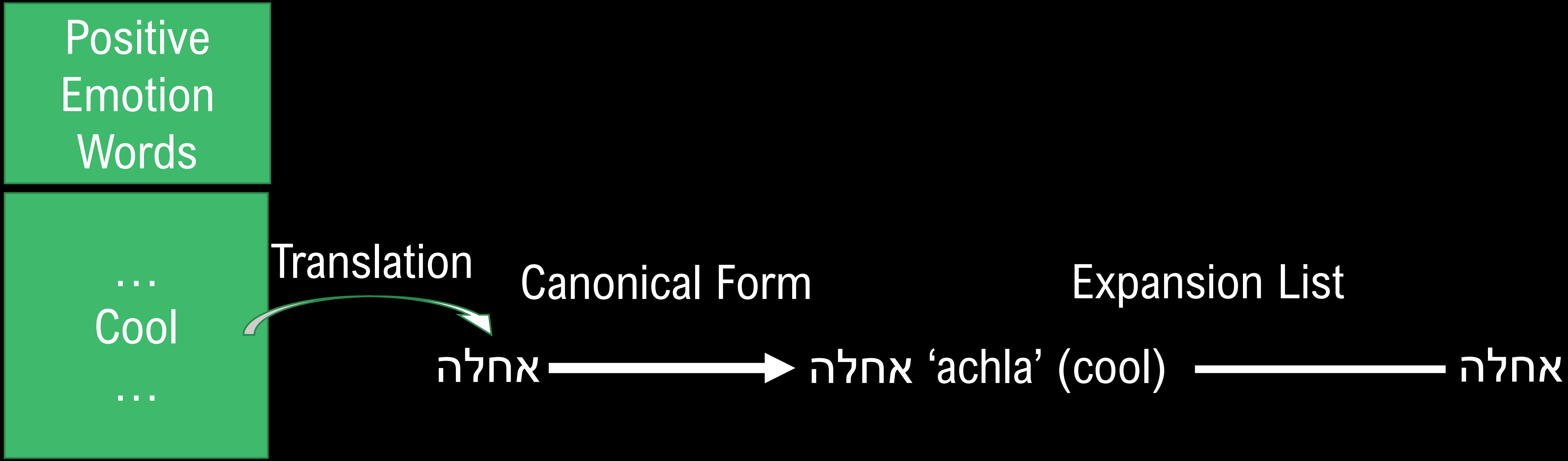
חימה ‘chema’ (anger)

Non-emotion word

חמה ‘chama’ (sun)

חמה

Solution



Positive Emotion Words

…

Cool

…

Translation

Canonical Form

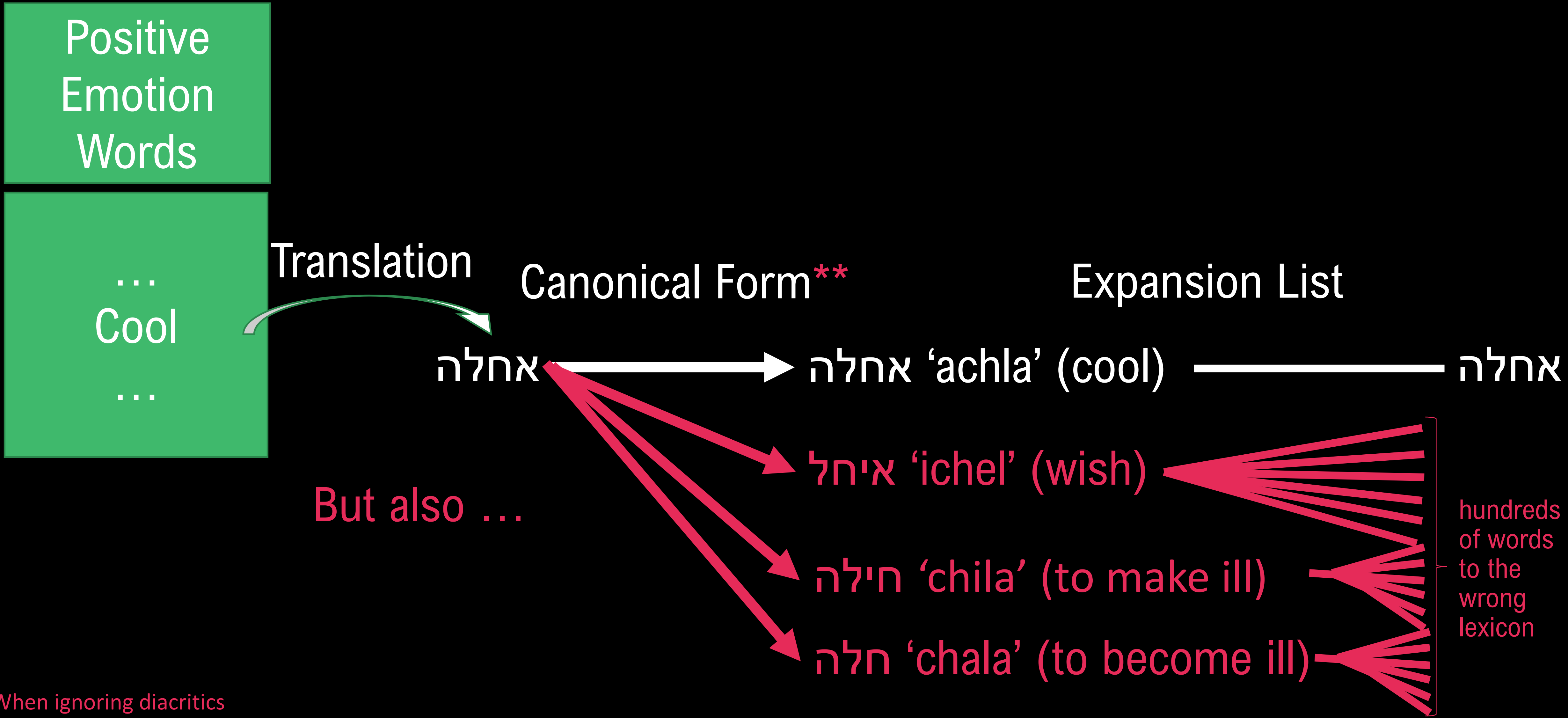
Expansion List

אחלה

אחלה ‘achla’ (cool)

אחלה

Solution…?



Positive Emotion Words

…
Cool
…

Translation

Canonical Form**

Expansion List

אחלה

אחלה ‘achla’ (cool)

אחלה

But also …

איחל ‘ichel’ (wish)

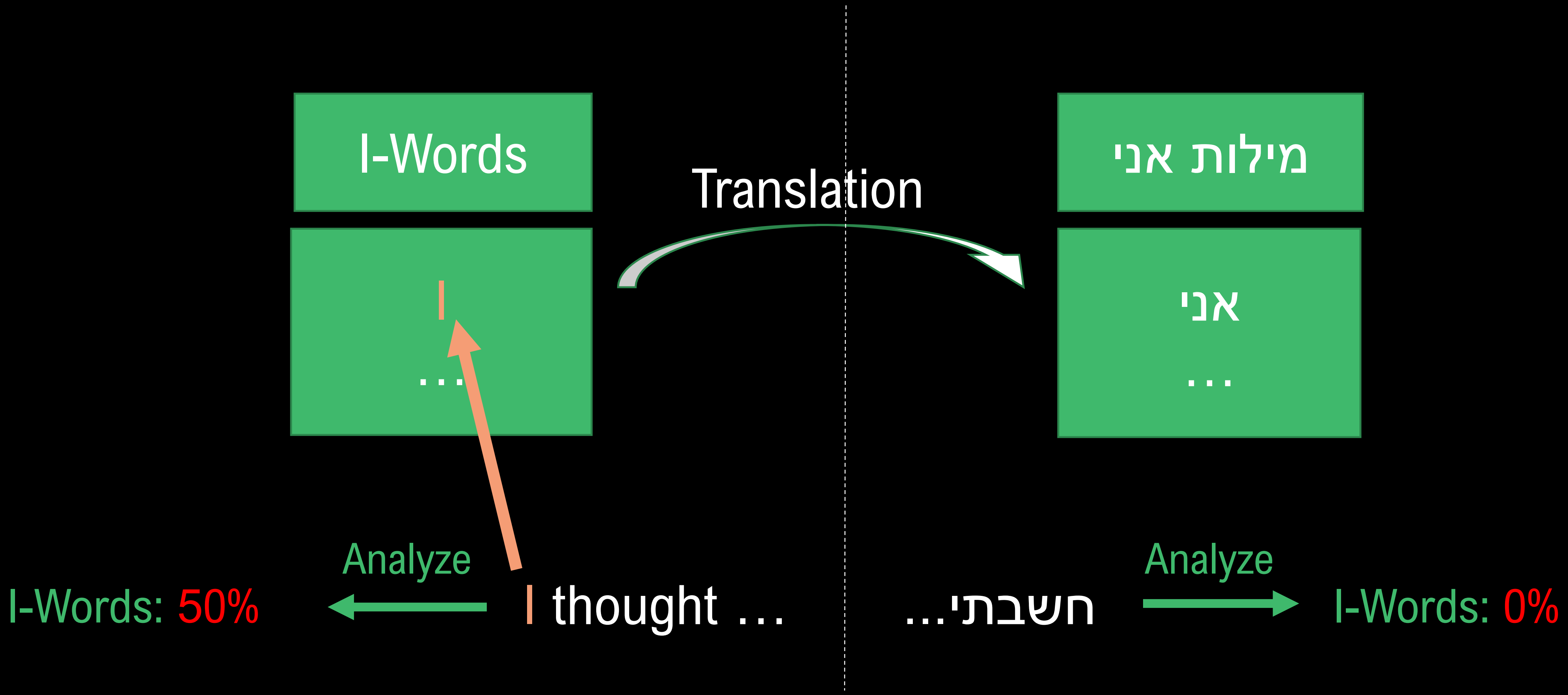
חילה ‘chila’ (to make ill)

חלה ‘chala’ (to become ill)

hundreds of words to the wrong lexicon

** When ignoring diacritics

Morphology
(3) Content expressed in morphology and not in words



I-Words

I

…

Translation

מילות אני

אני

…

Analyze

I-Words: 50%

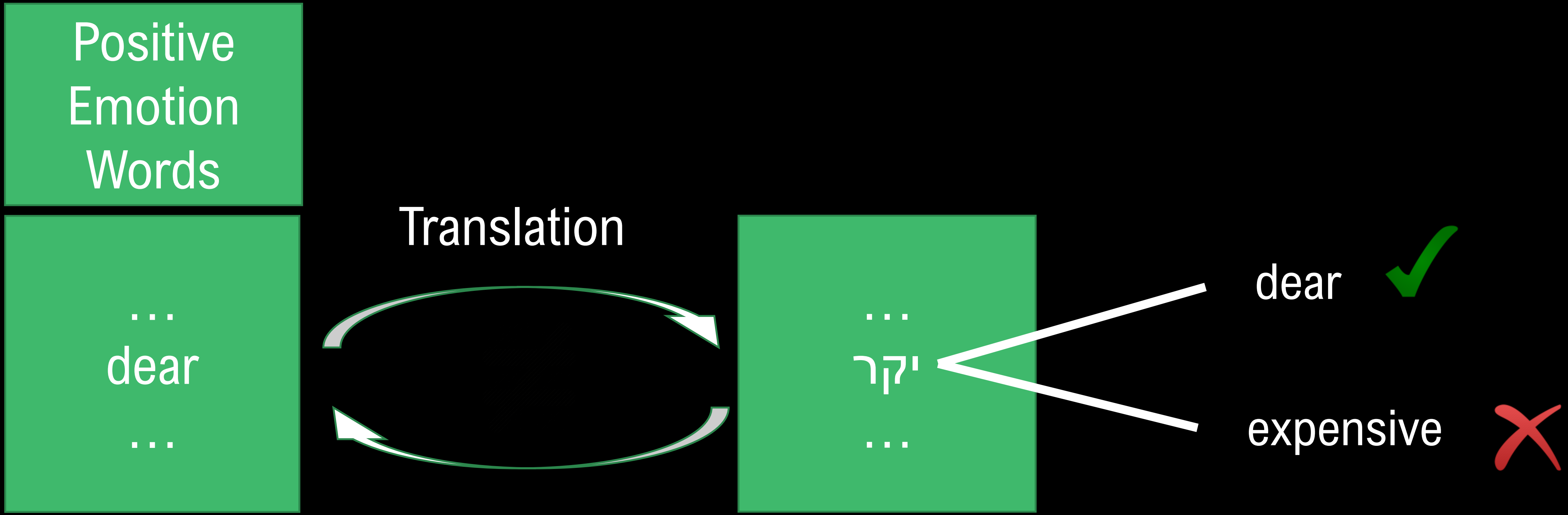
I thought …

חשבתי…

Analyze

I-Words: 0%

Ambiguity



Positive Emotion Words

…
dear
…

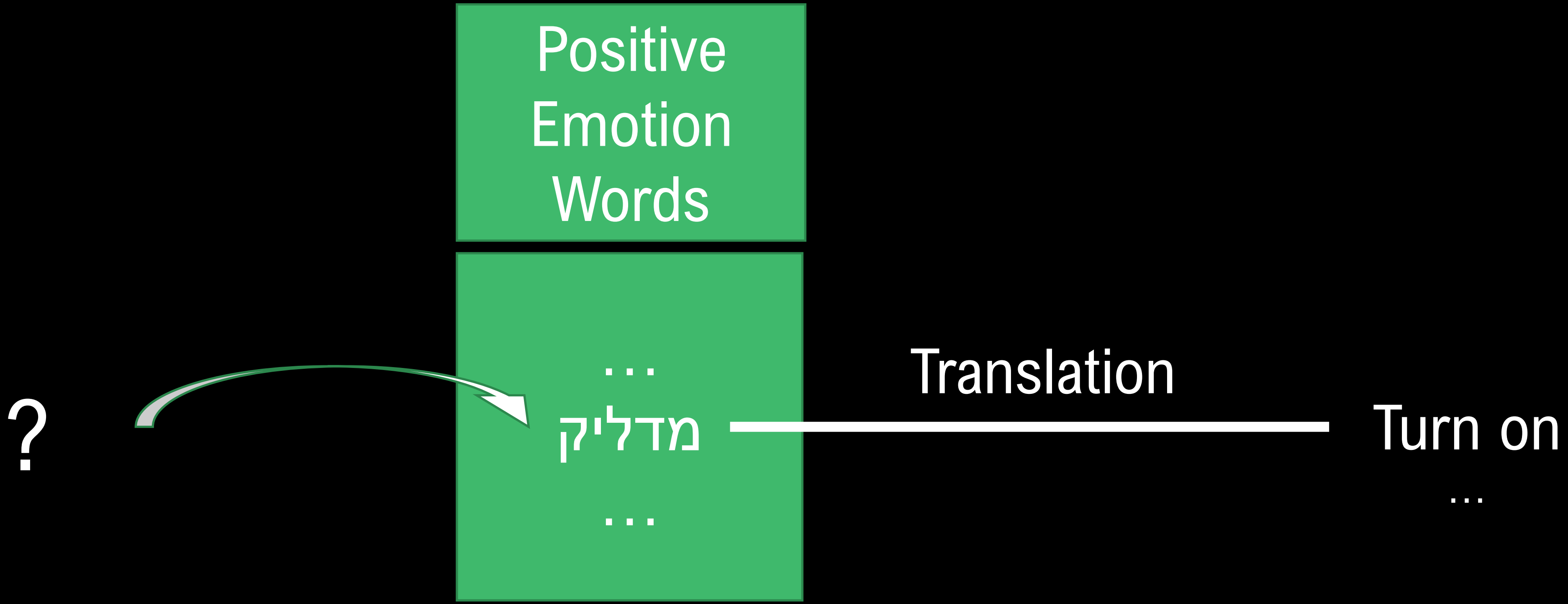
Translation

…
יקר
…

dear ✓

expensive ✗

Lack of Corresponding One Word



Positive Emotion Words

…

מדליק

…

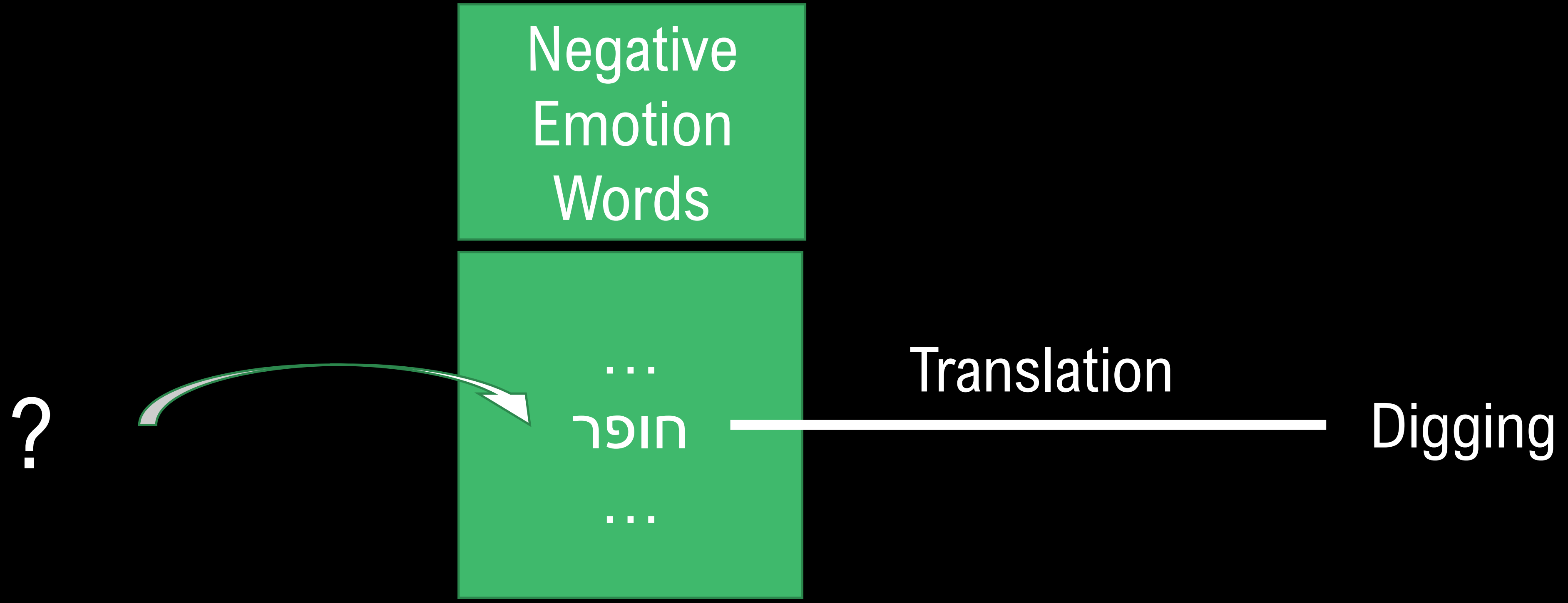
?

Translation

Turn on

…

Culture



Negative Emotion Words

…

חופר

…

?

Translation

Digging

Methodological Considerations in Lexicon Construction Process

Base Dataset

11.57 AVG (SD=3.18) Transcriptions X 74 Dyads = 872 Transcriptions

Statistics	Total	Per Patient	Per Session
Talk Turns	~150K	~2K	~200
Client Tokens	~4 Million	~50K	~4.5K
All Tokens	~5 Million	~70K	~6K

Client content is 80% of the session

Psychotherapy Research Lab
Bar-Ilan University

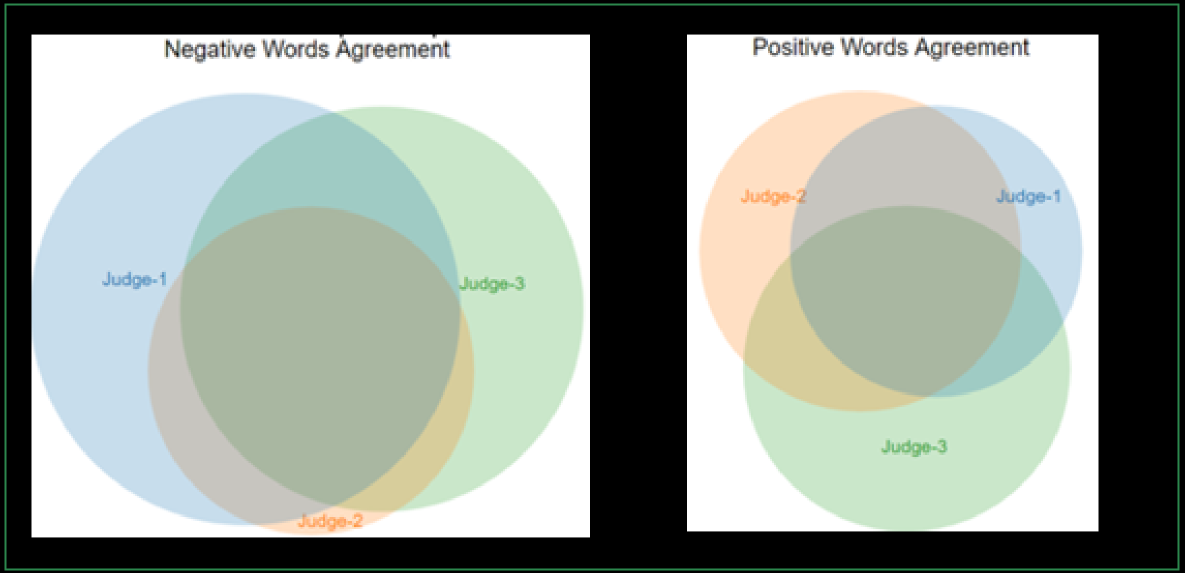
Monitoring Psychotherapy

Prof. Rivka Tuval-Mashiach, Dr. Dana Atzil-Slonim; Israel Science Foundation (grants 1348/15)

Lexicons Based on Expert Knowledge

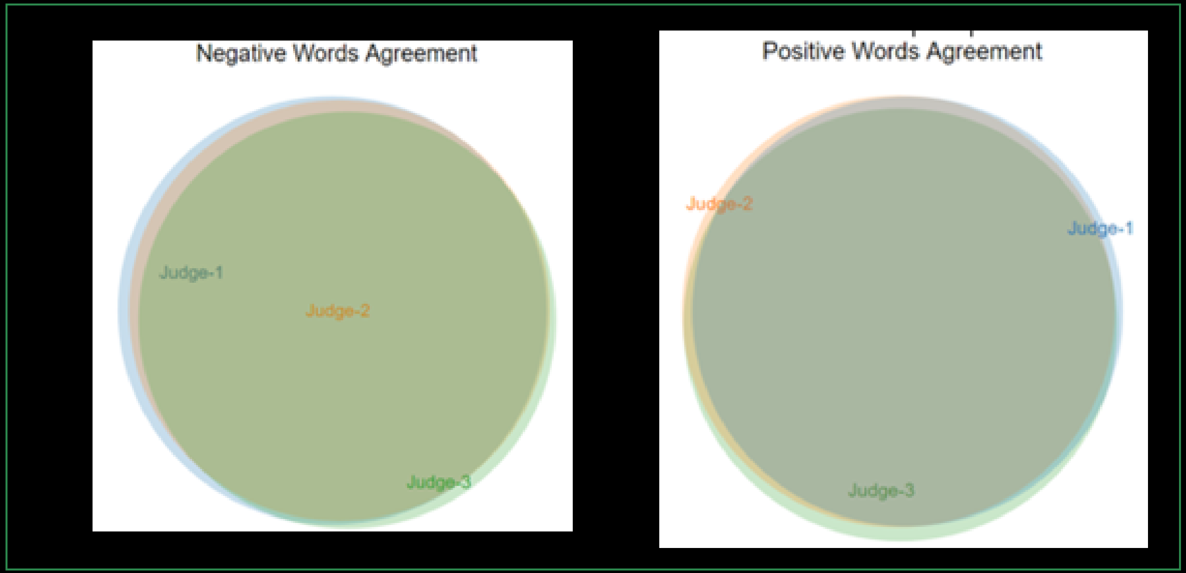
Valance (Positive & Negative)

Before Reconciliation Process



Fleiss' Kappa 0.54
** Moderate agreement

After Reconciliation Process



Fleiss' Kappa 0.95
** Almost perfect agreement

Valance
(Positive & Negative)

- Coverage
  - 2000 most frequent words cover 86% of all tokens in all transcripts
  - 15% of the all tokens in the transcripts were emotion words

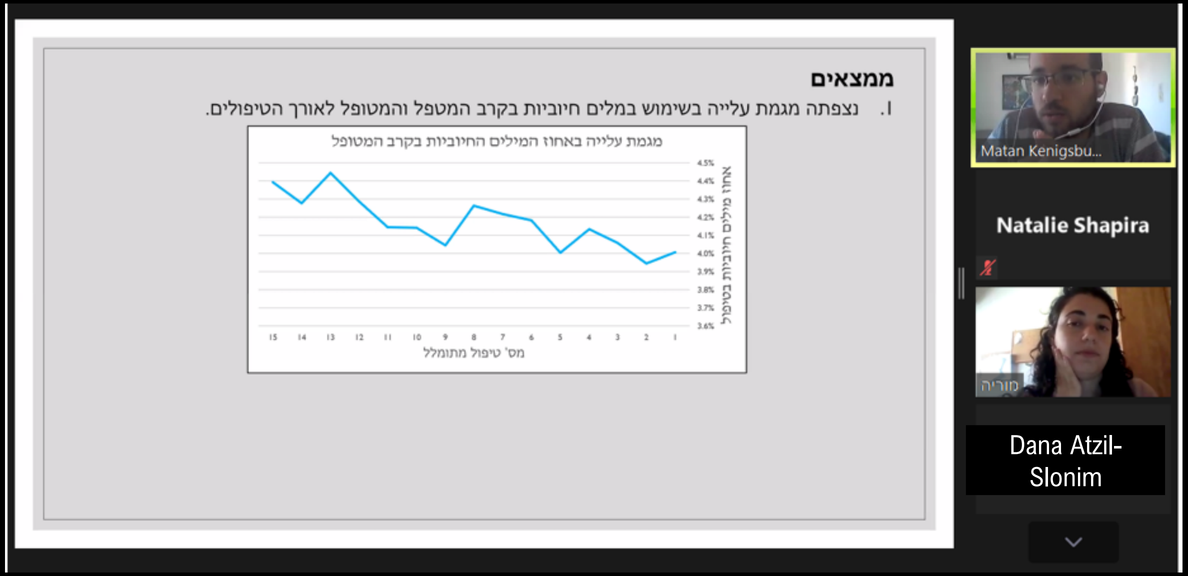
Valance (Positive & Negative)

Changes in emotion words associate with change in clients' functioning from session to session

↑ Negative Emotion Words ↓ Positive Emotion Words $\propto$ ↓ ORS From Session to Session

Shapira Natalie, Gal Lazarus, Yoav Goldberg, Eva Gilboa-Schechtman, Rivka Tuval-Mashiach, Daniel Juravski, and Dana Atzil-Slonim. "Using computerized text analysis to examine associations between linguistic features and clients' distress during psychotherapy." *Journal of counseling psychology* (2020).

Valance (Positive & Negative)



ממצאים

ו. נצפתה מגמת עלייה בשימוש במלים חיוביות בקרב המטפל והמטופל לאורך הטיפולים.

Emotion words associate with emotions

Negative Emotion Words   Positive Emotion Words   ∝   Client's and therapist's positive/negative emotions as reported in the POMS questionnaire

Rubin Moria, Kenigsbuch Matan, Shapira Natalie and Dana Atzil-Slonim. "Correlation between Emotion Words and the Emotional Experience in Psychological Therapy", Dept. of Psychology, Bar-Ilan University (2020)

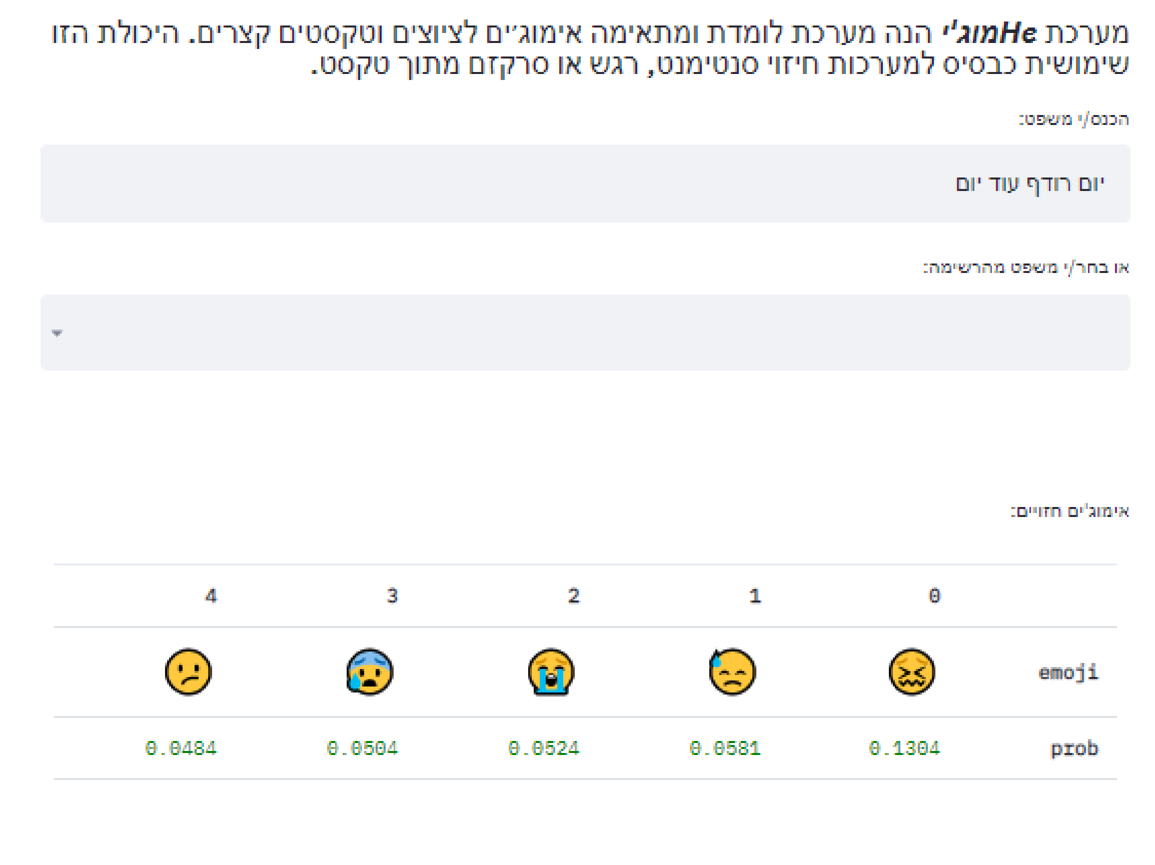
Valance (Positive & Negative)

Negative Emotion Words

Positive Emotion Words

∝

Predicted emojis by a pretrained model based on Twitter data



מערכת ***Heמוג׳י*** הנה מערכת לומדת ומתאימה אימוג׳ים לציוצים וטקסטים קצרים. היכולת הזו
שימושית כבסיס למערכות חיזוי סנטימנט, רגש או סרקזם מתוך טקסט.

הכנס/י משפט:

יום רודף עוד יום

או בחר/י משפט מהרשימה:

אימוג׳ים חזויים:

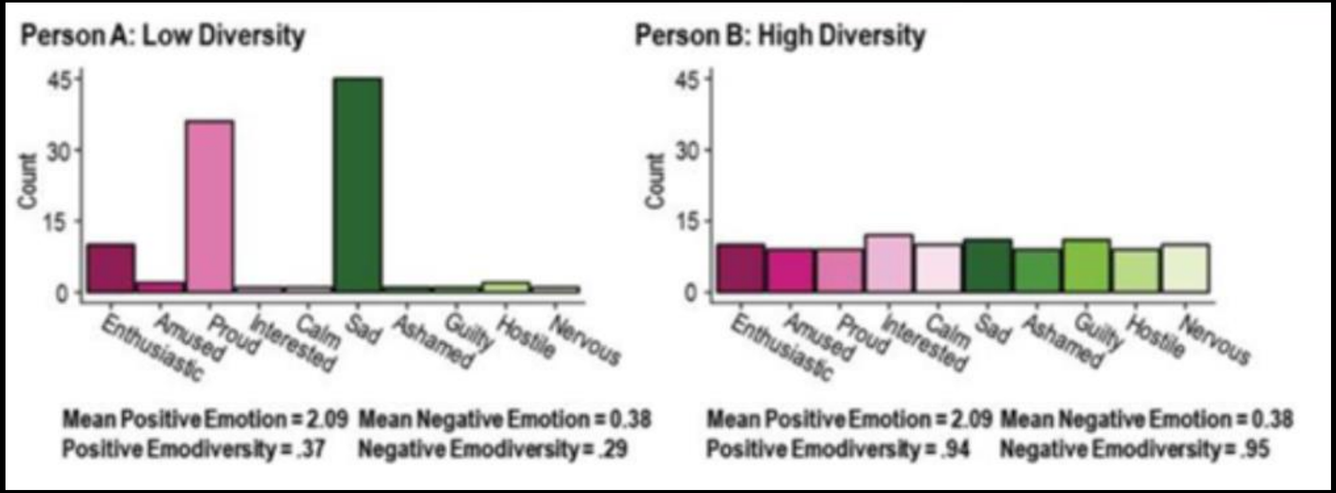
4	3	2	1	0	
😕	😰	😭	😓	😖	emoji
0.0484	0.0504	0.0524	0.0581	0.1304	prob

https://hub.docker.com/r/danieljuravski/hemoji

Juravski Daniel, Natural Language Processing Methods for Analyzing Textual Psychotherapy Data, Under the supervision of Yoav Goldberg. Dept. of Computer Science, Bar-Ilan University (2020)

Emotional Variety

Motivation:



Ong et al. (2018)

Emotional Variety

- Enthusiastic
- Amused
- Proud
- Interested
- Calm
- Sad
- Ashamed
- Guilty
- Hostile
- Nervous
- Anger
- Contentment
- Anxiety
- Vigor
- Joy
- Disgust
- Surprise
- Trust
- Anticipation
- Confusion
- Fatigue

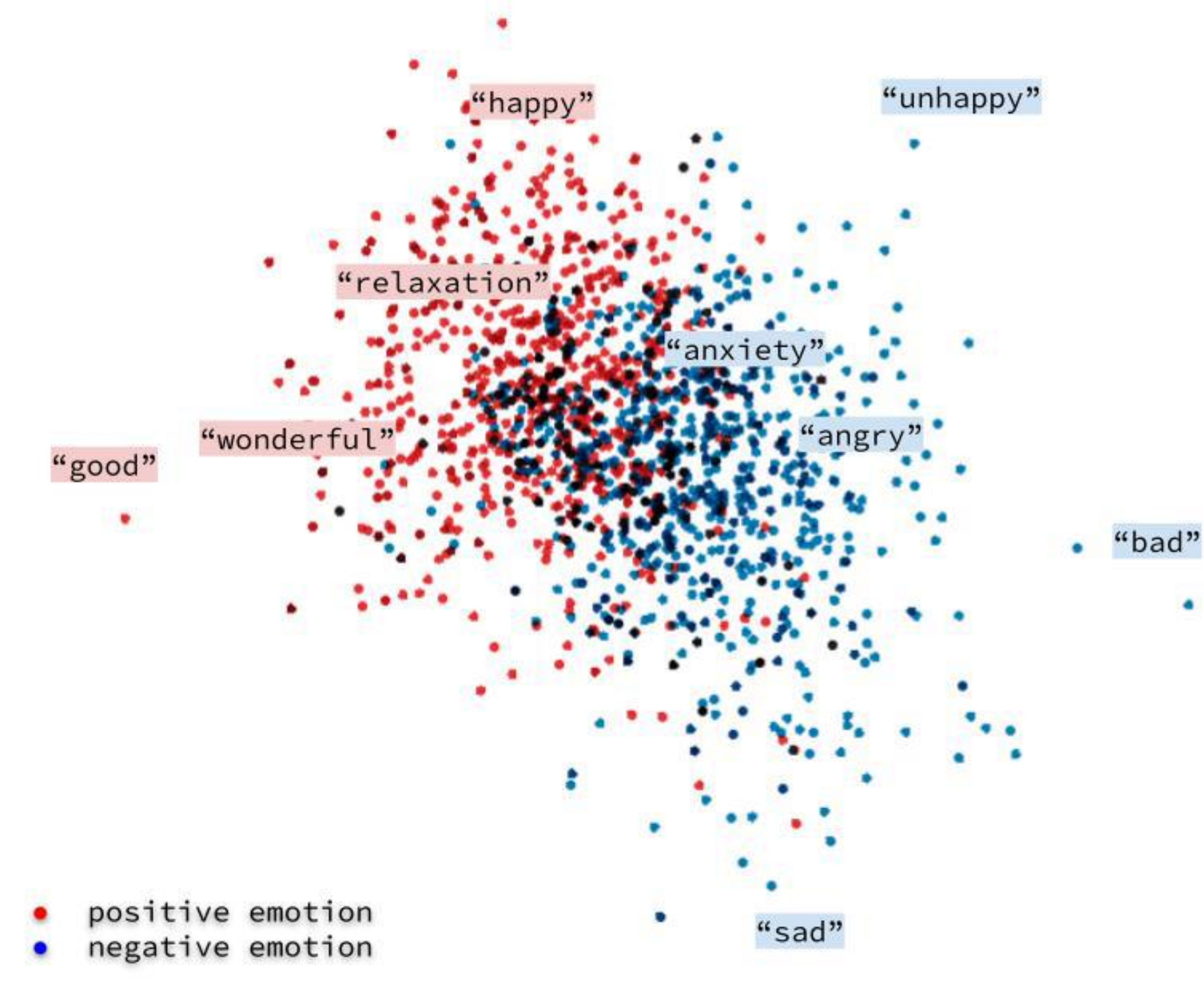
List composed based on the POMS emotion questionnaire, Robert plutchik's "wheel of emotions" (Plutchik, 2000) and emotions described by Ong et al. (2018).

Emotional Variety

Complementing-emotion

Enthusiastic	Not Enthusiastic	Guilty	Not Guilty	Joy	Not Joy
Amused	Not Amused	Hostile	Not Hostile	Disgust	Not Disgust
Proud	Not Proud	Nervous	Not Nervous	Surprise	Not Surprise
Interested	Not Interested	Anger	Not Anger	Trust	Not Trust
Calm	Not Calm	Contentment	Not Contentment	Anticipation	Not Anticipation
Sad	Not Sad	Anxiety	Not Anxiety	Confusion	Not Confusion
Ashamed	Not Ashamed	Vigor	Not Vigor	Fatigue	Not Fatigue

2D-Projection of emotion word embeddings



Emotional Variety

- Freely-suggested words by 19 advanced undergraduate psychology students
- 5000 most frequent words, covering 90% of all tokens in all transcripts

Merged by majority of three judges

Emotional Variety

- Freely-suggested words by 19 advanced undergraduate psychology students
- 5000 most frequent words, covering 90% of all tokens in all transcripts
- Automatic seed expansion (under construction)

Merged by majority of three judges

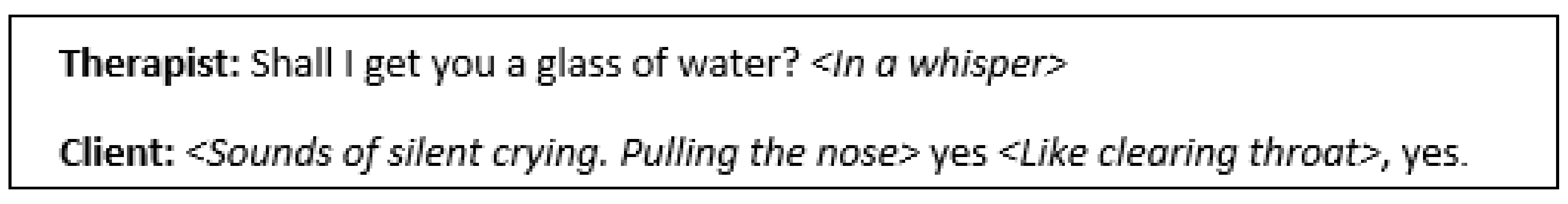
Emotional Variety

Research use case – work in progress

- Freely-suggested words by 19 advanced undergraduate psychology students
- 5000 most frequent words, covering 90% of all tokens in all transcripts
- Automatic seed expansion (under construction)

Merged by majority of three judges

Paralinguistics Events

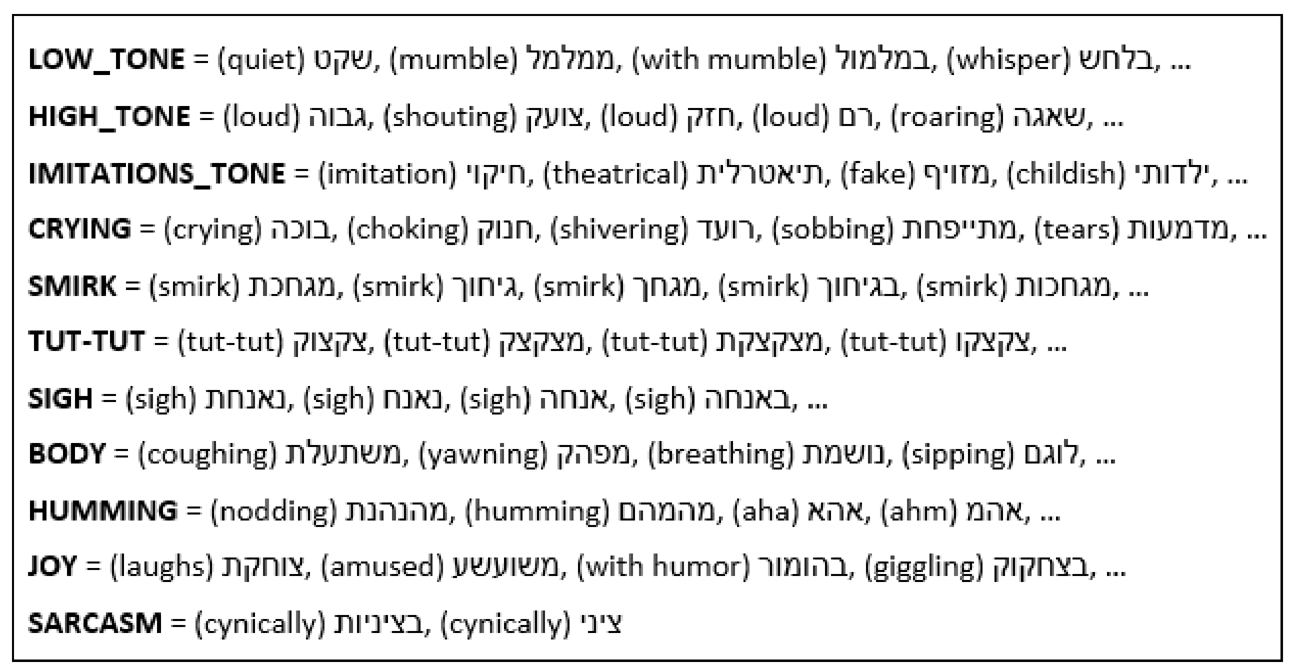


Therapist: Shall I get you a glass of water? *<In a whisper>*

Client: *<Sounds of silent crying. Pulling the nose>* yes *<Like clearing throat>*, yes.

Paralinguistics Events

- 1022 word-type appeared at least twice (out of 31,067 tokens)
- 11 categories characterized by domain experts
- Each word-type classified (100% agreement)



LOW_TONE = (quiet) שקט, (mumble) ממלמל, (with mumble) במלמול, (whisper) בלחש, ...

HIGH_TONE = (loud) גבוה, (shouting) צועק, (loud) חזק, (loud) רם, (roaring) שאגה, ...

IMITATIONS_TONE = (imitation) חיקוי, (theatrical) תיאטרלית, (fake) מזויף, (childish) ילדותי, ...

CRYING = (crying) בוכה, (choking) חנוק, (shivering) רועד, (sobbing) מתייפחת, (tears) מדמעות, ...

SMIRK = (smirk) מגחכת, (smirk) גיחוך, (smirk) מגחך, (smirk) בגיחוך, (smirk) מגחכות, ...

TUT-TUT = (tut-tut) צקצוק, (tut-tut) מצקצק, (tut-tut) מצקצקת, (tut-tut) צקצקו, ...

SIGH = (sigh) נאנחת, (sigh) נאנח, (sigh) אנחה, (sigh) באנחה, ...

BODY = (coughing) משתעלת, (yawning) מפהק, (breathing) נושמת, (sipping) לוגם, ...

HUMMING = (nodding) מהנהנת, (humming) מהמהם, (aha) אהא, (ahm) אהמ, ...

JOY = (laughs) צוחקת, (amused) משועשע, (with humor) בהומור, (giggling) בצחקוק, ...

SARCASM = (cynically) בציניות, (cynically) ציני

Paralinguistics Events

I. Negative Emotion Words Positive Emotion Words $\propto$ Paralinguistics Events

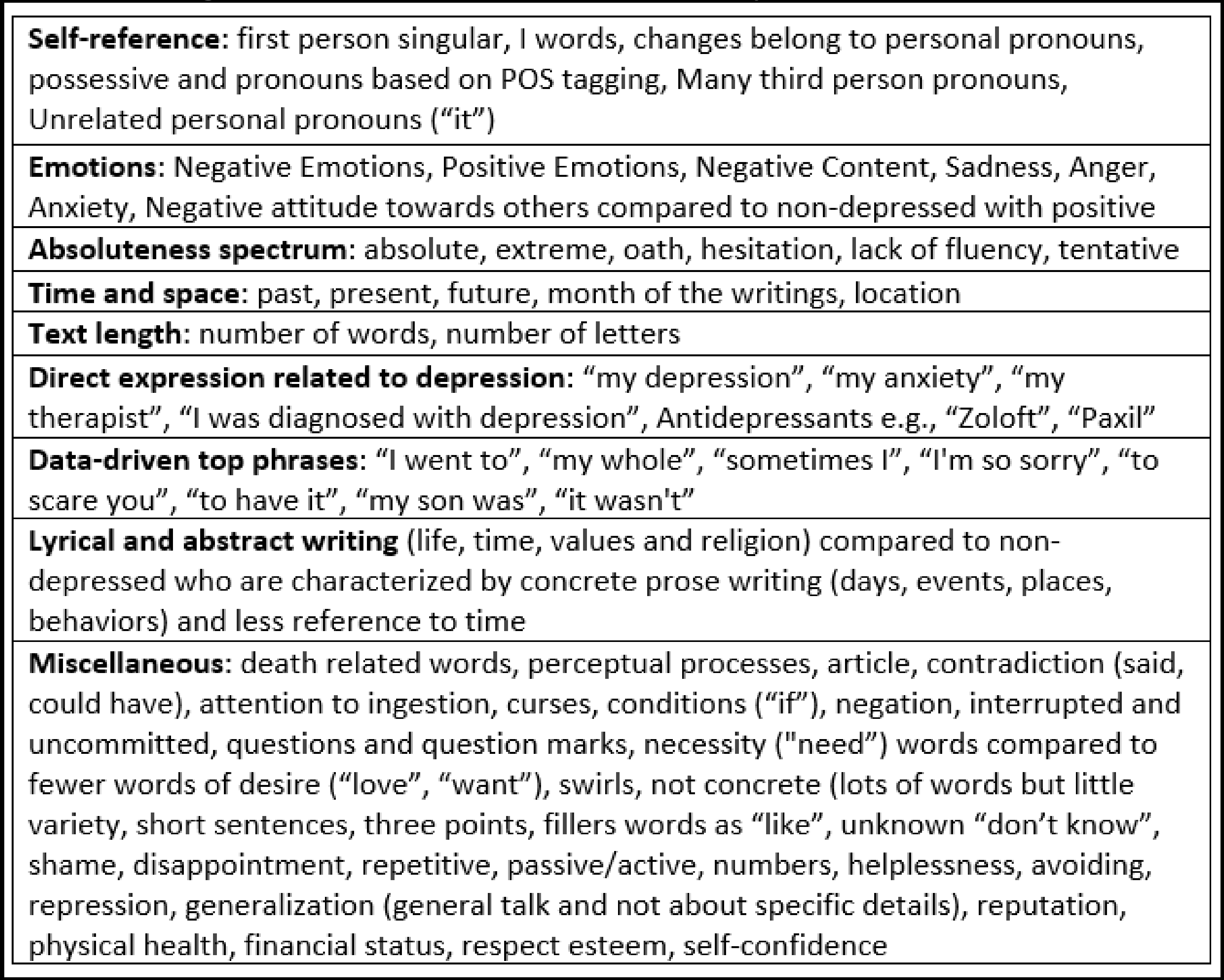
II. Therapist Paralinguistics Events $\propto$ Client Paralinguistics Events

Nadaf Adam,Yosef Noam, Shapira Natalie and Dana Atzil-Slonim. “Synchrony in paralinguistics events”
Dept. of Psychology, Bar-Ilan University (2020)

Depressive Characteristics

- Literature review
- Case studies

Linguistic Characteristics of Depressive Texts



Self-reference: first person singular, I words, changes belong to personal pronouns, possessive and pronouns based on POS tagging, Many third person pronouns, Unrelated personal pronouns (“it”)

Emotions: Negative Emotions, Positive Emotions, Negative Content, Sadness, Anger, Anxiety, Negative attitude towards others compared to non-depressed with positive

Absoluteness spectrum: absolute, extreme, oath, hesitation, lack of fluency, tentative

Time and space: past, present, future, month of the writings, location

Text length: number of words, number of letters

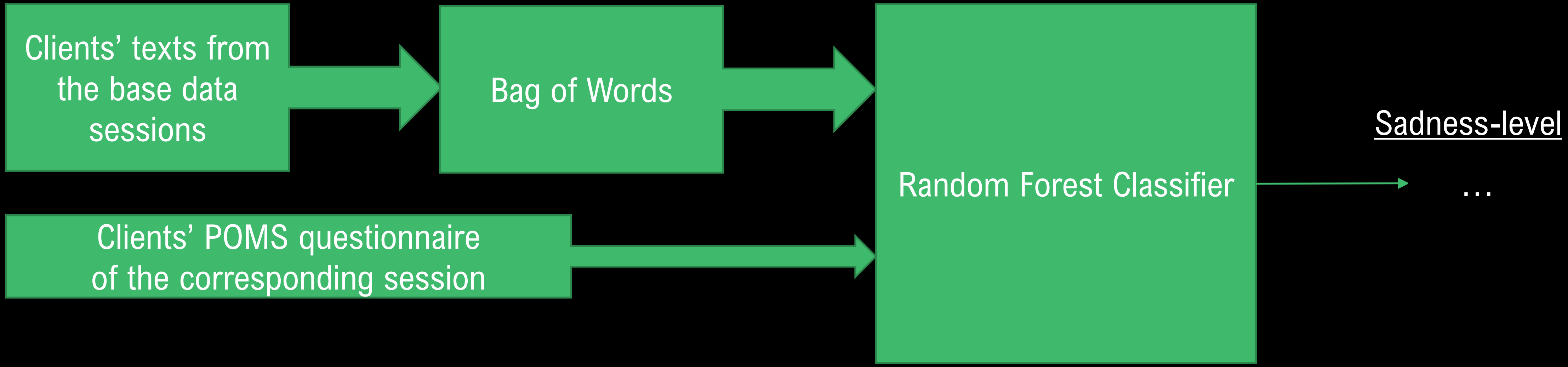
Direct expression related to depression: “my depression”, “my anxiety”, “my therapist”, “I was diagnosed with depression”, Antidepressants e.g., “Zoloft”, “Paxil”

Data-driven top phrases: “I went to”, “my whole”, “sometimes I”, “I'm so sorry”, “to scare you”, “to have it”, “my son was”, “it wasn't”

Lyrical and abstract writing (life, time, values and religion) compared to non-depressed who are characterized by concrete prose writing (days, events, places, behaviors) and less reference to time

Miscellaneous: death related words, perceptual processes, article, contradiction (said, could have), attention to ingestion, curses, conditions (“if”), negation, interrupted and uncommitted, questions and question marks, necessity ("need”) words compared to fewer words of desire (“love”, “want”), swirls, not concrete (lots of words but little variety, short sentences, three points, fillers words as “like”, unknown “don’t know”, shame, disappointment, repetitive, passive/active, numbers, helplessness, avoiding, repression, generalization (general talk and not about specific details), reputation, physical health, financial status, respect esteem, self-confidence

Depressive Characteristics



Clients' texts from the base data sessions

Bag of Words

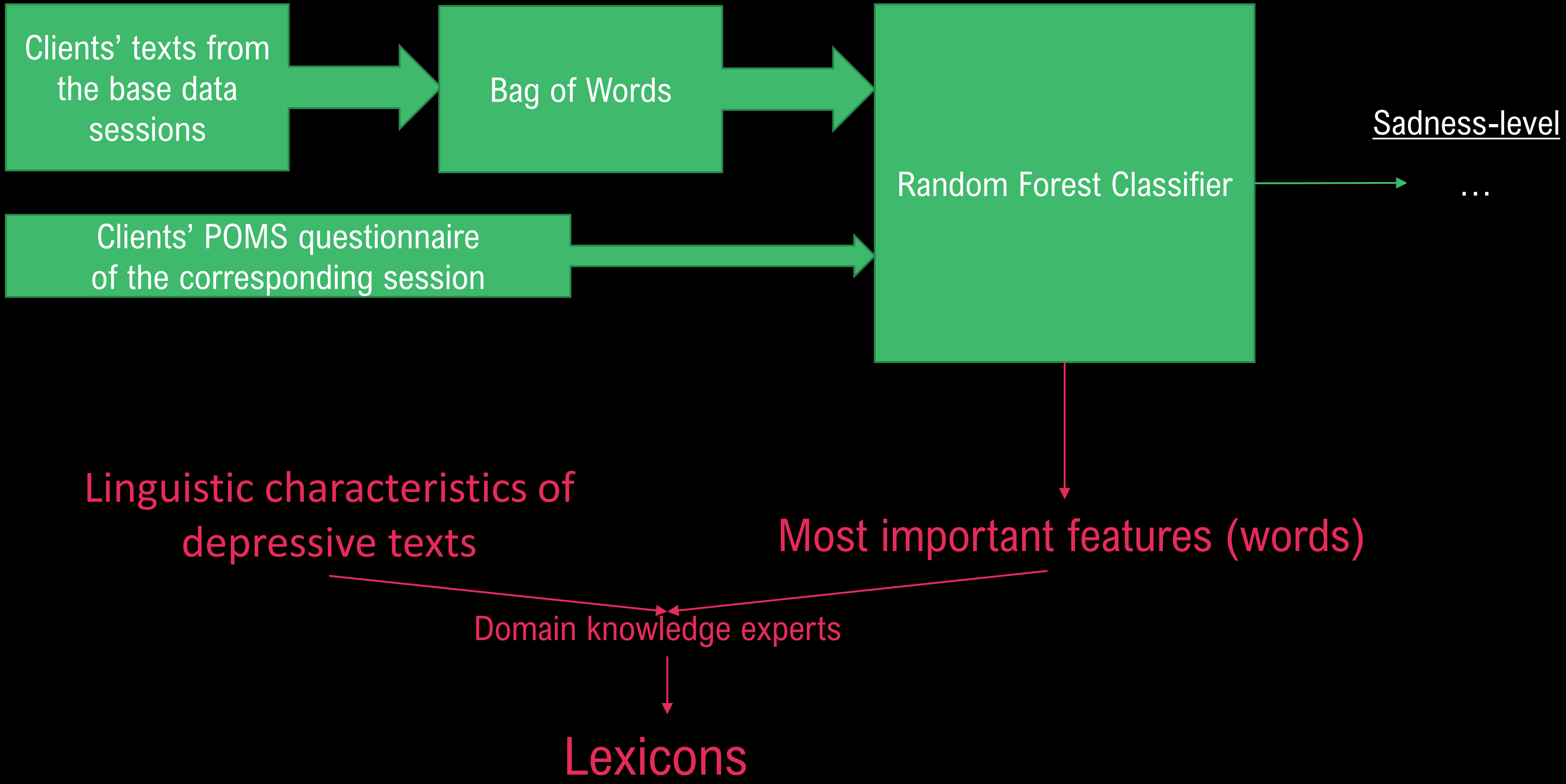
Clients' POMS questionnaire of the corresponding session

Random Forest Classifier

Sadness-level

…

Depressive Characteristics



Clients' texts from the base data sessions

Bag of Words

Clients' POMS questionnaire of the corresponding session

Random Forest Classifier

Sadness-level

…

Linguistic characteristics of depressive texts

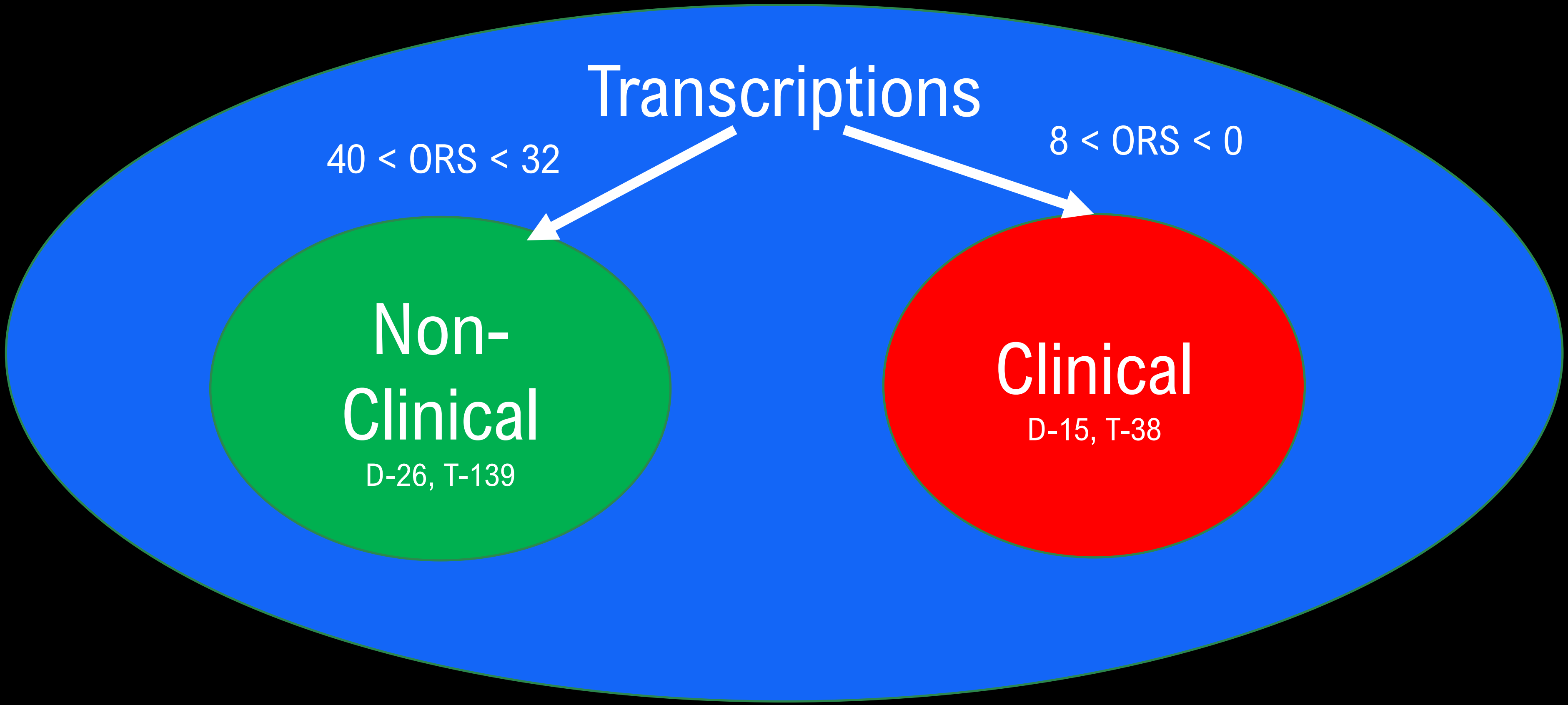
Most important features (words)

Domain knowledge experts

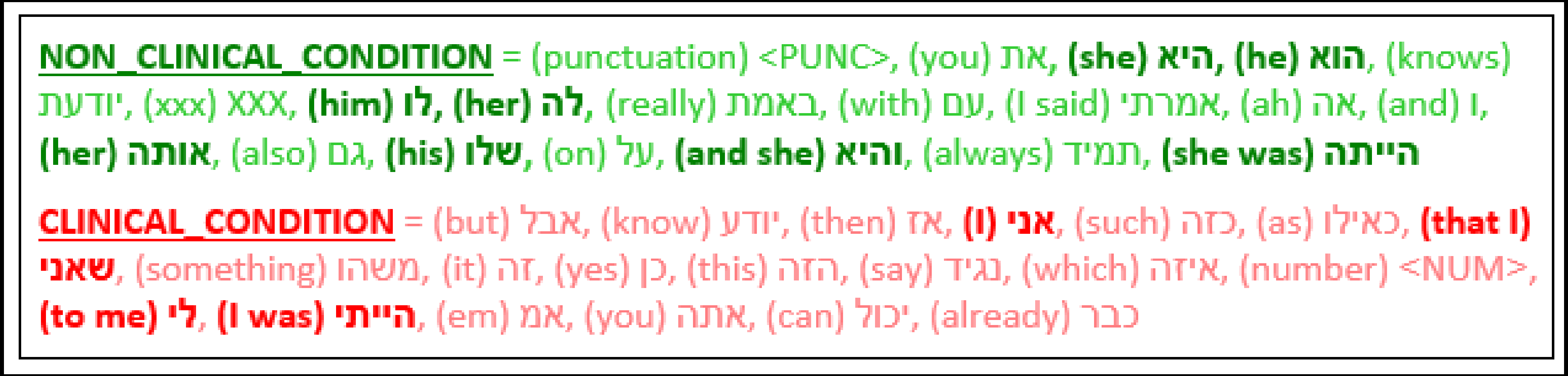
Lexicons

Data-Driven Word Lists

Well-Being


Transcriptions
40 < ORS < 32
8 < ORS < 0
Non-Clinical
D-26, T-139
Clinical
D-15, T-38

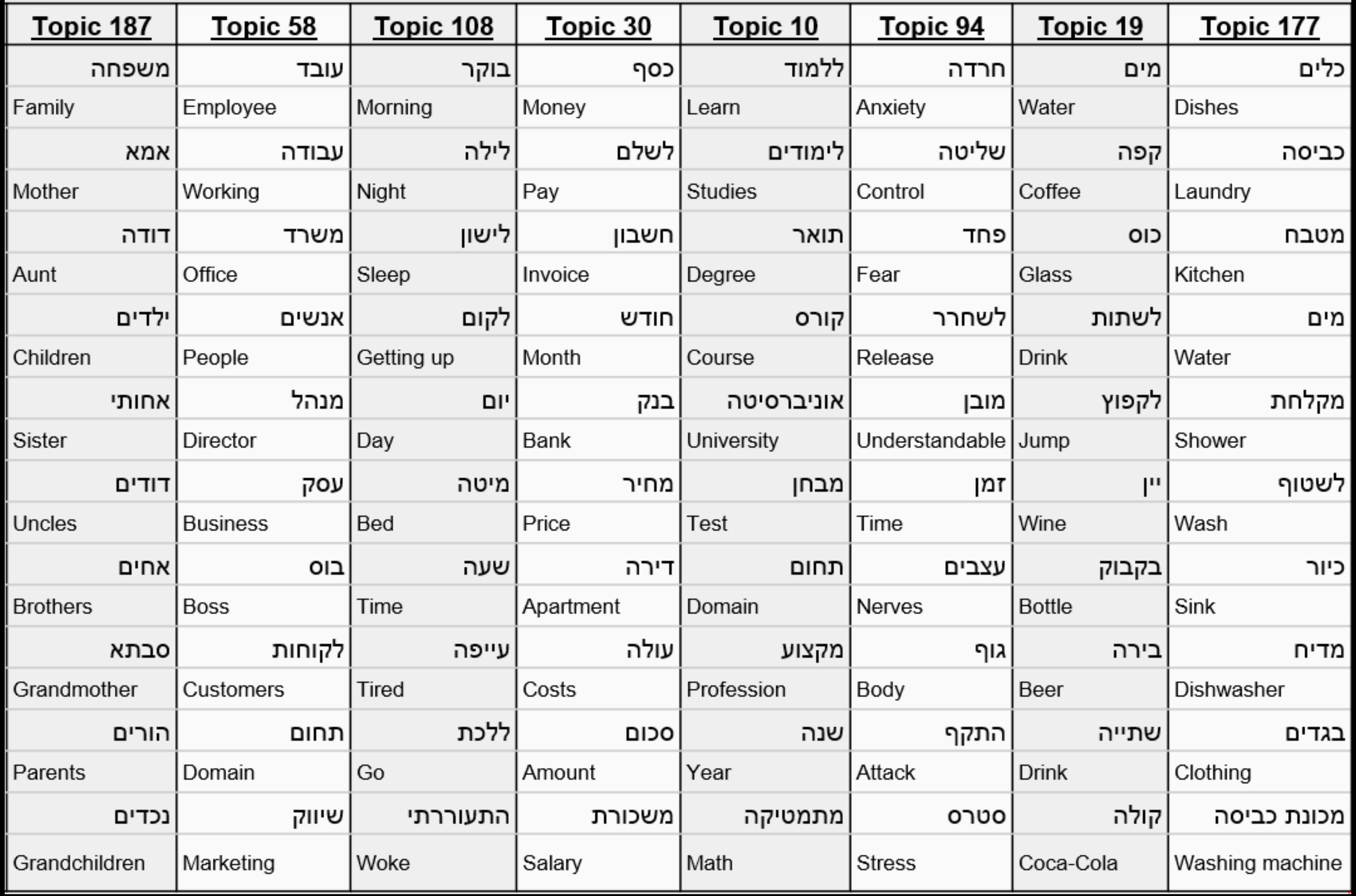
Well-Being



NON_CLINICAL_CONDITION = (punctuation) <PUNC>, (you) את, **(she) היא, (he) הוא**, (knows) יודעת, (xxx) XXX, **(him) לו, (her) לה**, (really) באמת, (with) עם, (I said) אמרתי, (ah) אה, (and) ו, **(her) אותה**, (also) גם, **(his) שלו**, (on) על, **(and she) והיא**, (always) תמיד, **(she was) הייתה**

CLINICAL_CONDITION = (but) אבל, (know) יודע, (then) אז, **(I) אני**, (such) כזה, (as) כאילו, **(that I) שאני**, (something) משהו, (it) זה, (yes) כן, (this) הזה, (say) נגיד, (which) איזה, (number) <NUM>, **(to me) לי**, **(I was) הייתי**, (em) אמ, (you) אתה, (can) יכול, (already) כבר

Conversation Topics



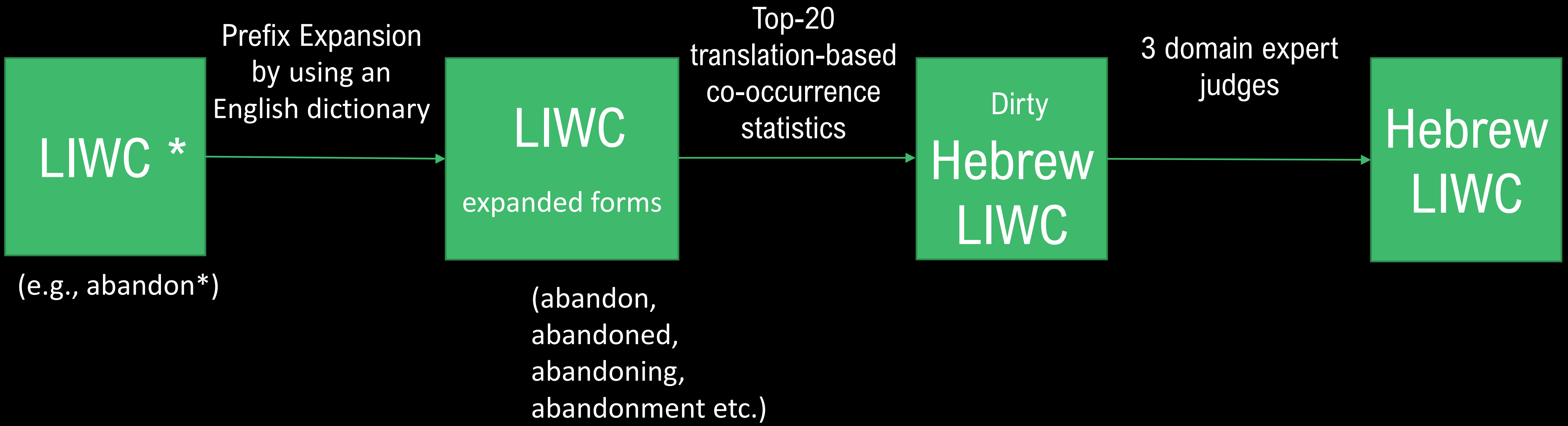
Topic 187	Topic 58	Topic 108	Topic 30	Topic 10	Topic 94	Topic 19	Topic 177
משפחה	עובד	בוקר	כסף	ללמוד	חרדה	מים	כלים
Family	Employee	Morning	Money	Learn	Anxiety	Water	Dishes
אמא	עבודה	לילה	לשלם	לימודים	שליטה	קפה	כביסה
Mother	Working	Night	Pay	Studies	Control	Coffee	Laundry
דודה	משרד	לישון	חשבון	תואר	פחד	כוס	מטבח
Aunt	Office	Sleep	Invoice	Degree	Fear	Glass	Kitchen
ילדים	אנשים	לקום	חודש	קורס	לשחרר	לשתות	מים
Children	People	Getting up	Month	Course	Release	Drink	Water
אחותי	מנהל	יום	בנק	אוניברסיטה	מובן	לקפוץ	מקלחת
Sister	Director	Day	Bank	University	Understandable	Jump	Shower
דודים	עסק	מיטה	מחיר	מבחן	זמן	יין	לשטוף
Uncles	Business	Bed	Price	Test	Time	Wine	Wash
אחים	בוס	שעה	דירה	תחום	עצבים	בקבוק	כיור
Brothers	Boss	Time	Apartment	Domain	Nerves	Bottle	Sink
סבתא	לקוחות	עייפה	עולה	מקצוע	גוף	בירה	מדיח
Grandmother	Customers	Tired	Costs	Profession	Body	Beer	Dishwasher
הורים	תחום	ללכת	סכום	שנה	התקף	שתייה	בגדים
Parents	Domain	Go	Amount	Year	Attack	Drink	Clothing
נכדים	שיווק	התעוררתי	משכורת	מתמטיקה	סטרס	קולה	מכונת כביסה
Grandchildren	Marketing	Woke	Salary	Math	Stress	Coca-Cola	Washing machine

Atzil-Slonim Dana, Daniel Juravski, Eran Bar-Kalifa, Eva Gilboa-Schechtman, Rivka Tuval-Mashiach, Natalie Shapira, and Yoav Goldberg. "Using topic models to identify clients' functioning levels and alliance ruptures in psychotherapy." *Psychotherapy* (2021).

Lexicons Based on Expert Knowledge and Automatic Methods

(under construction)

Hebrew LIWC



LIWC *

(e.g., abandon*)

Prefix Expansion by using an English dictionary

LIWC

expanded forms

(abandon,
abandoned,
abandoning,
abandonment etc.)

Top-20 translation-based co-occurrence statistics

Dirty

Hebrew LIWC

3 domain expert judges

Hebrew LIWC

* Some of the categories are difficult or even impossible to translate into Hebrew.
For example, the *articles* lexicon (e.g., “a”, “an”, “the”, etc.) has no Hebrew equivalent nor does the *I words* lexicon

(under construction)

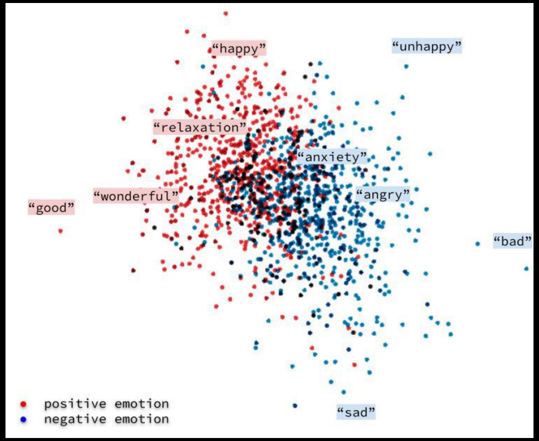
Expansions

Seed Lexicons

Sad

Expansions

(under construction)



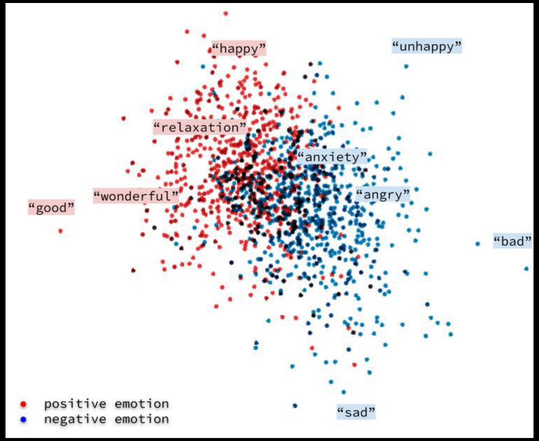
Seed Lexicons

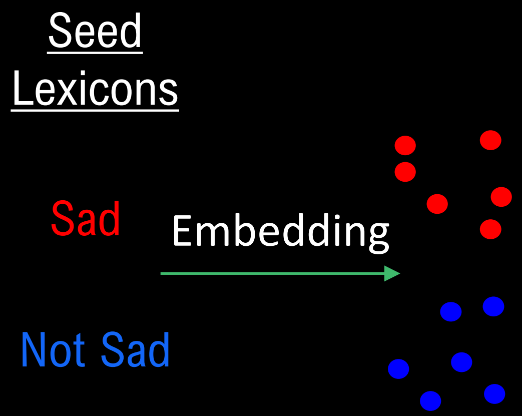
Sad

Not Sad

Expansions

(under construction)





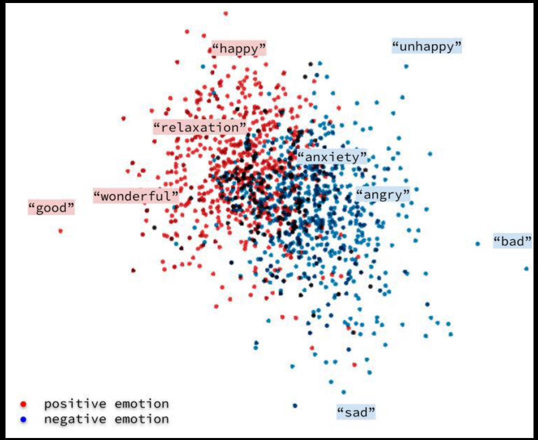
Seed Lexicons

Sad

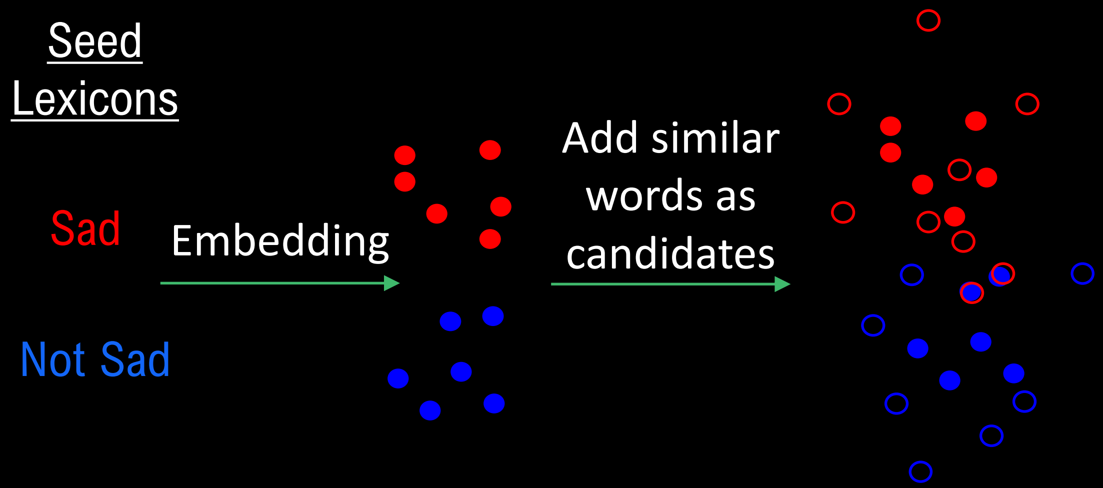
Embedding

Not Sad

(under construction)



Expansions


Seed Lexicons

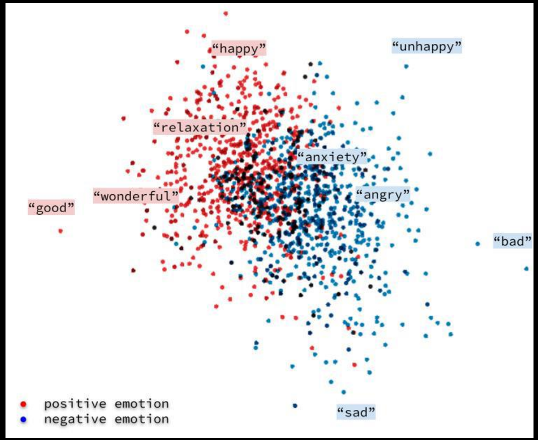
Sad

Not Sad

Embedding

Add similar words as candidates

(under construction)



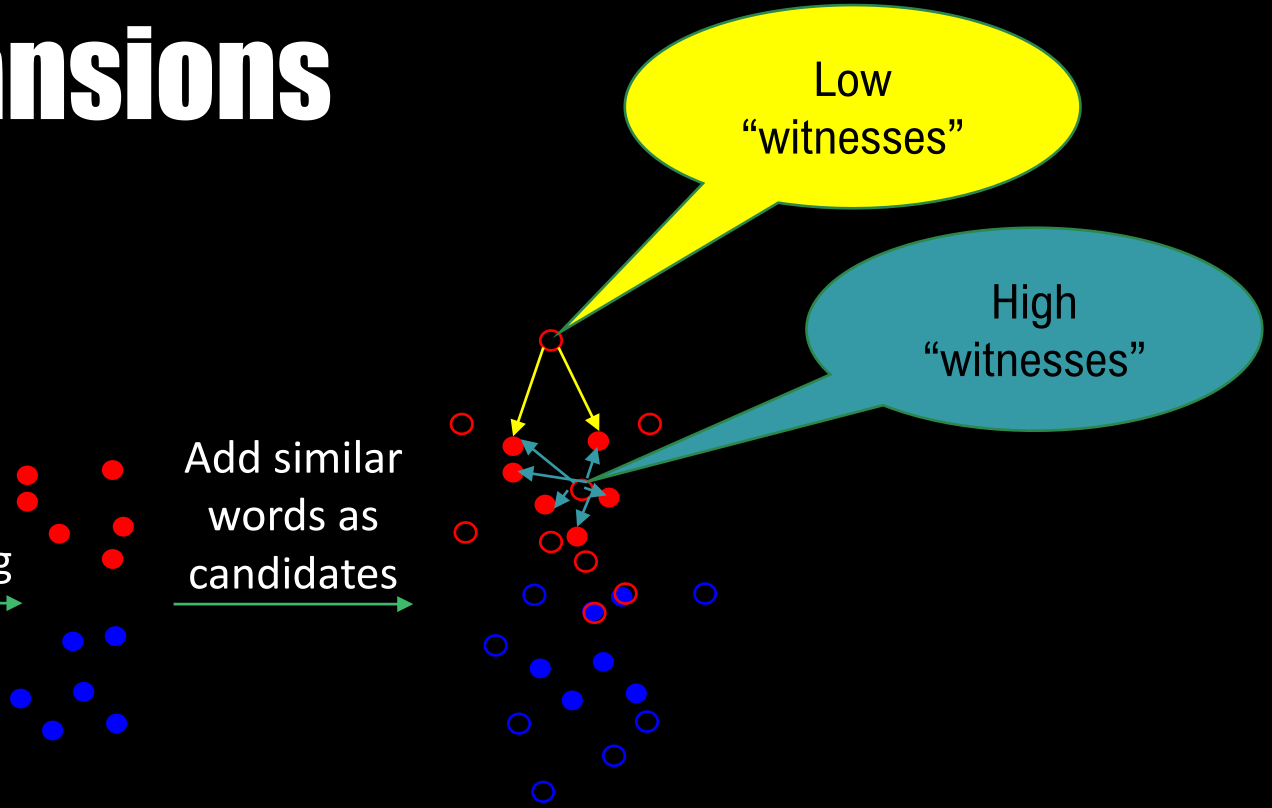
Expansions

Seed Lexicons

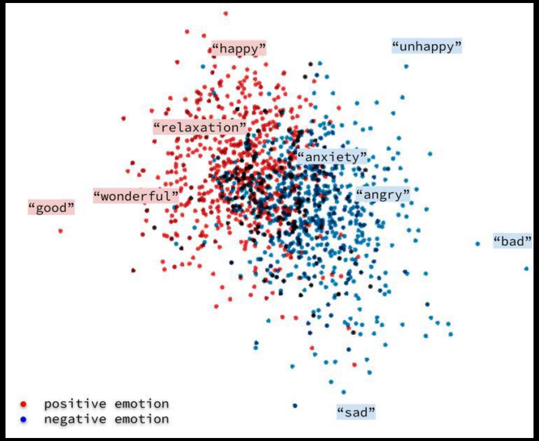
Sad

Not Sad

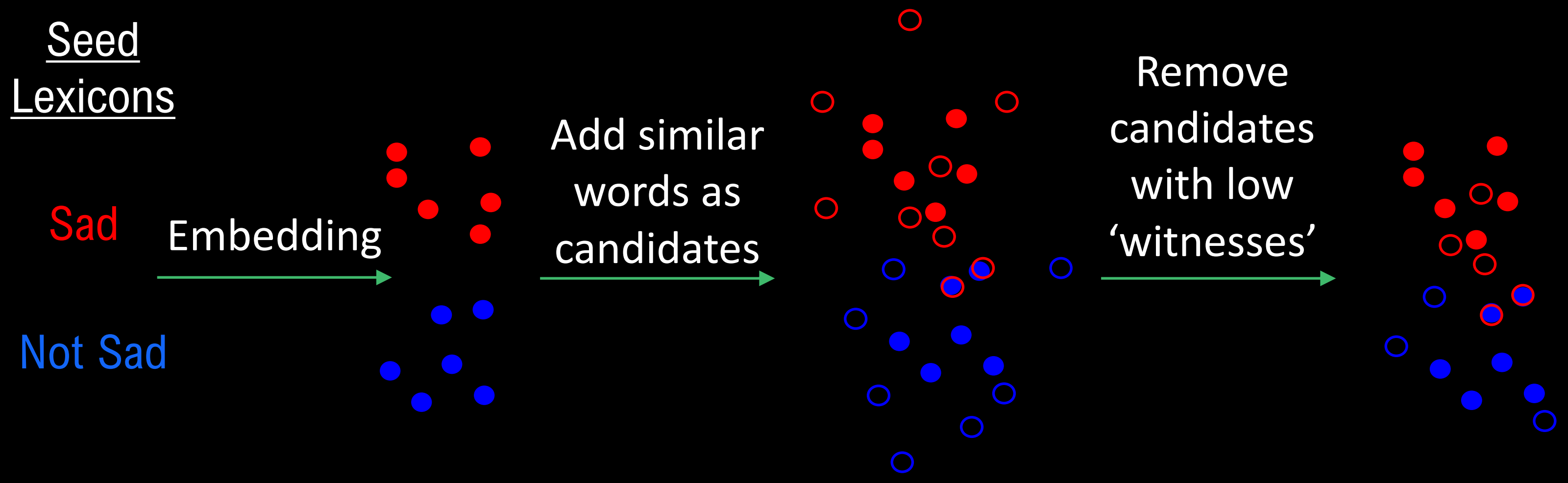
Embedding



(under construction)



Expansions


Seed Lexicons

Sad

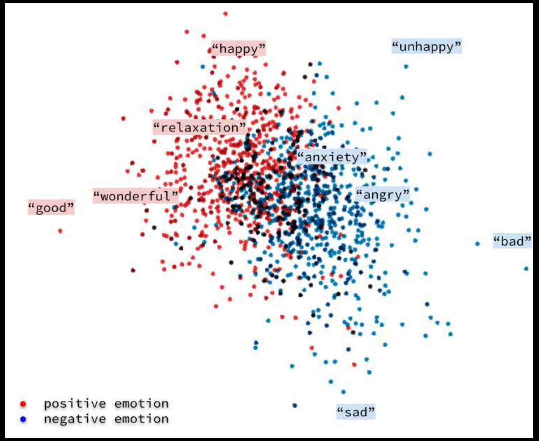
Not Sad

Embedding

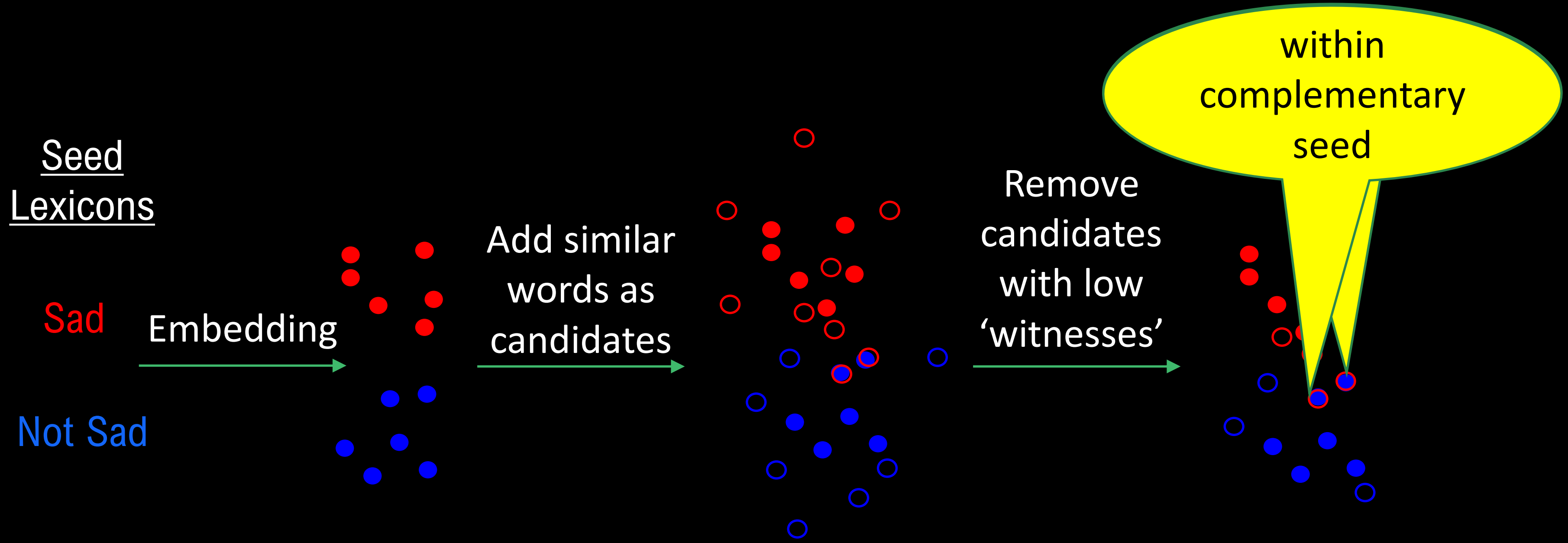
Add similar words as candidates

Remove candidates with low 'witnesses'

(under construction)



Expansions



Seed Lexicons

Sad

Not Sad

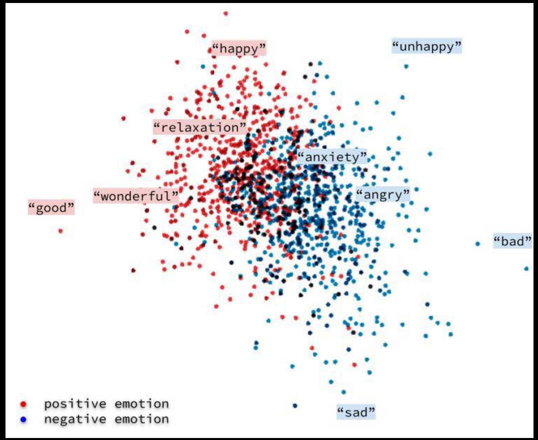
Embedding

Add similar words as candidates

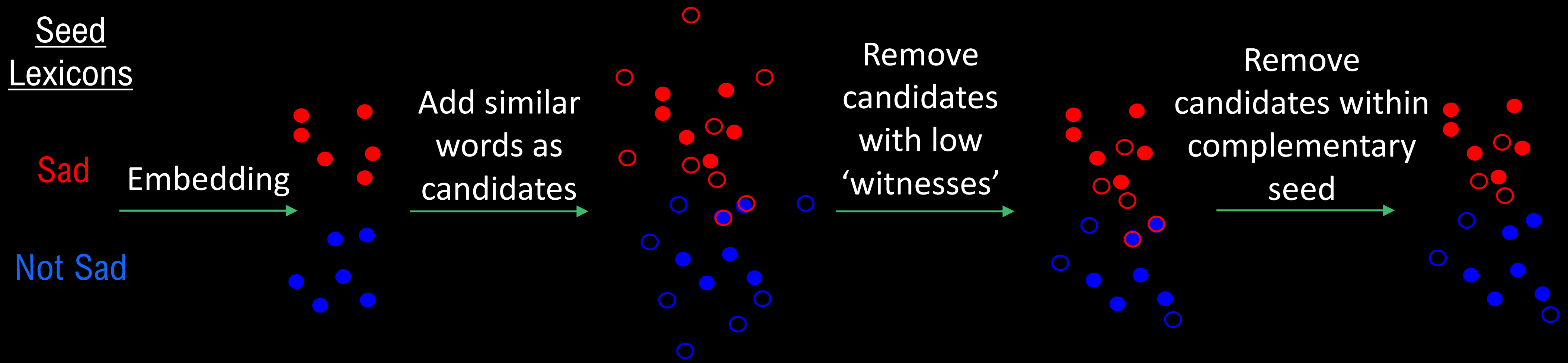
Remove candidates with low 'witnesses'

within complementary seed

(under construction)



Expansions



Seed Lexicons

Sad

Not Sad

Embedding

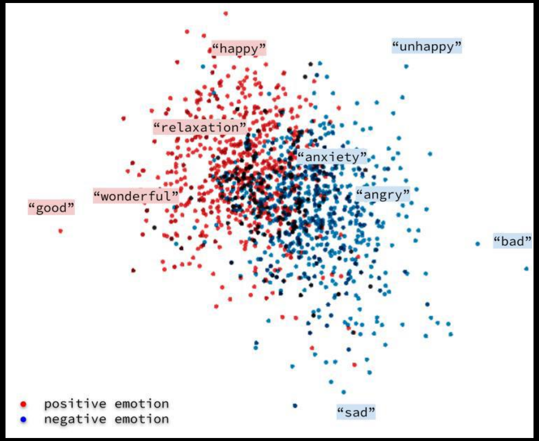
Add similar words as candidates

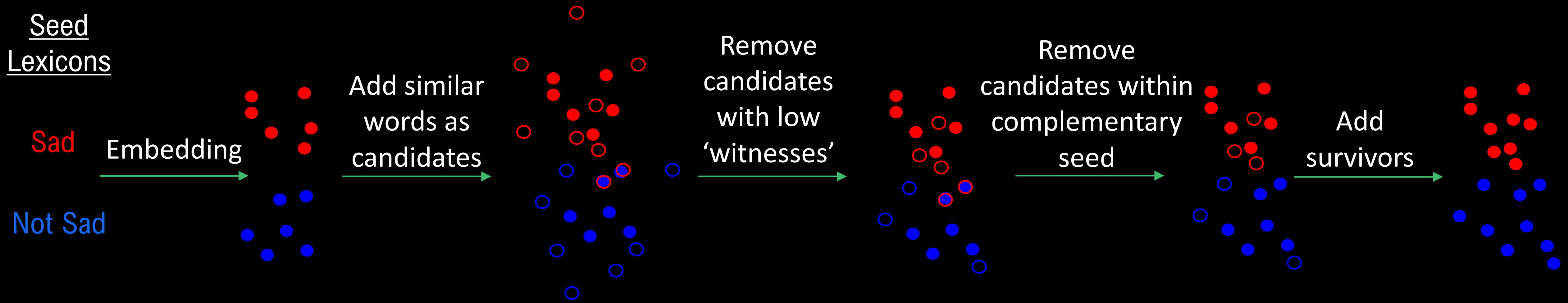
Remove candidates with low 'witnesses'

Remove candidates within complementary seed

Expansions

(under construction)





Seed Lexicons

Sad

Not Sad

Embedding

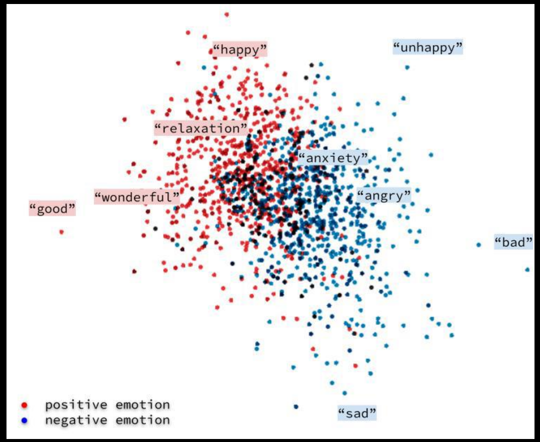
Add similar words as candidates

Remove candidates with low 'witnesses'

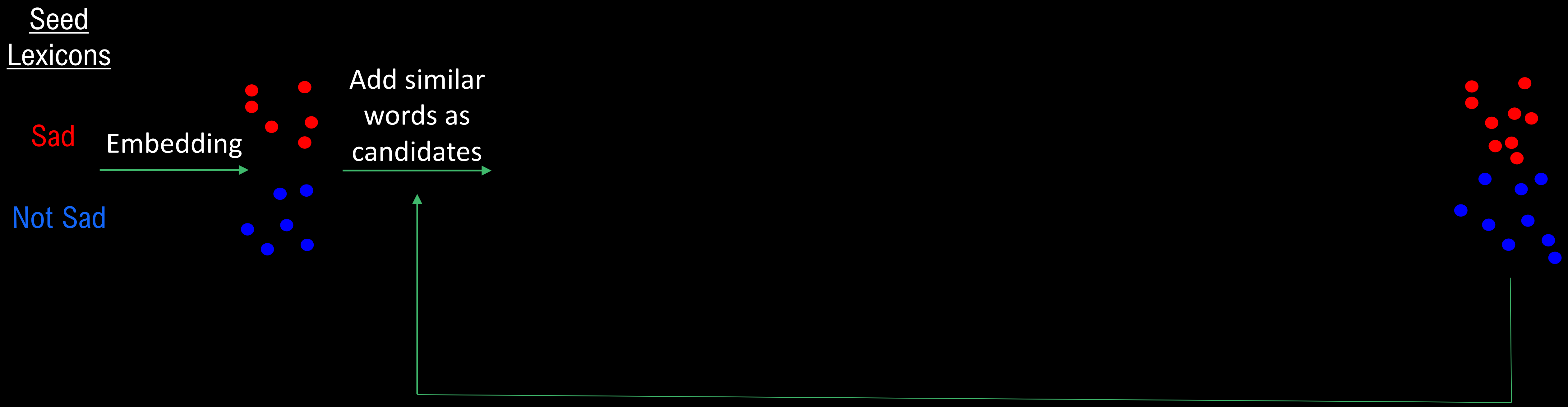
Remove candidates within complementary seed

Add survivors

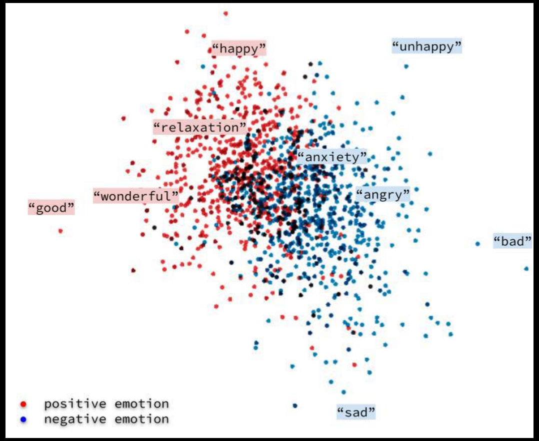
(under construction)



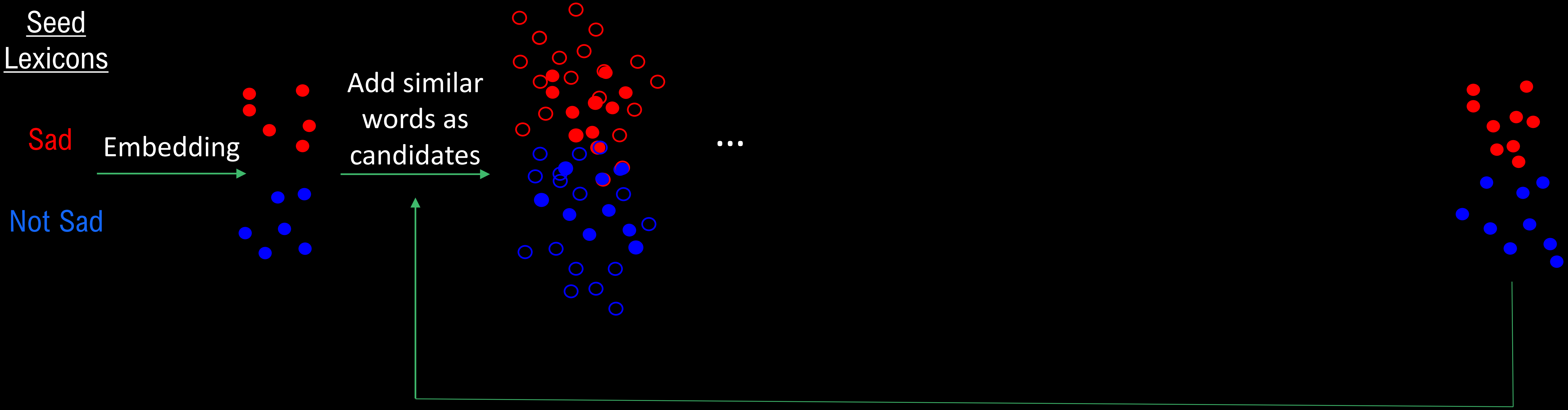
Expansions



(under construction)



Expansions



Seed Lexicons

Sad

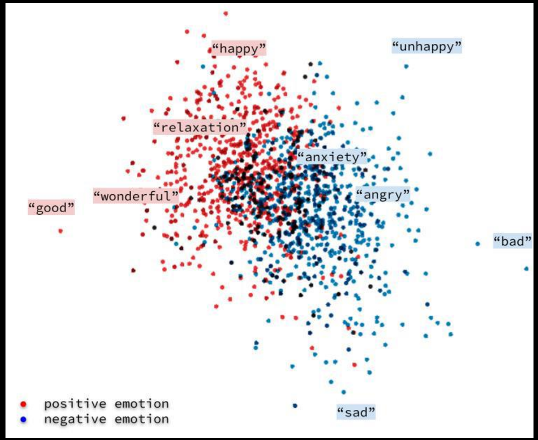
Not Sad

Embedding

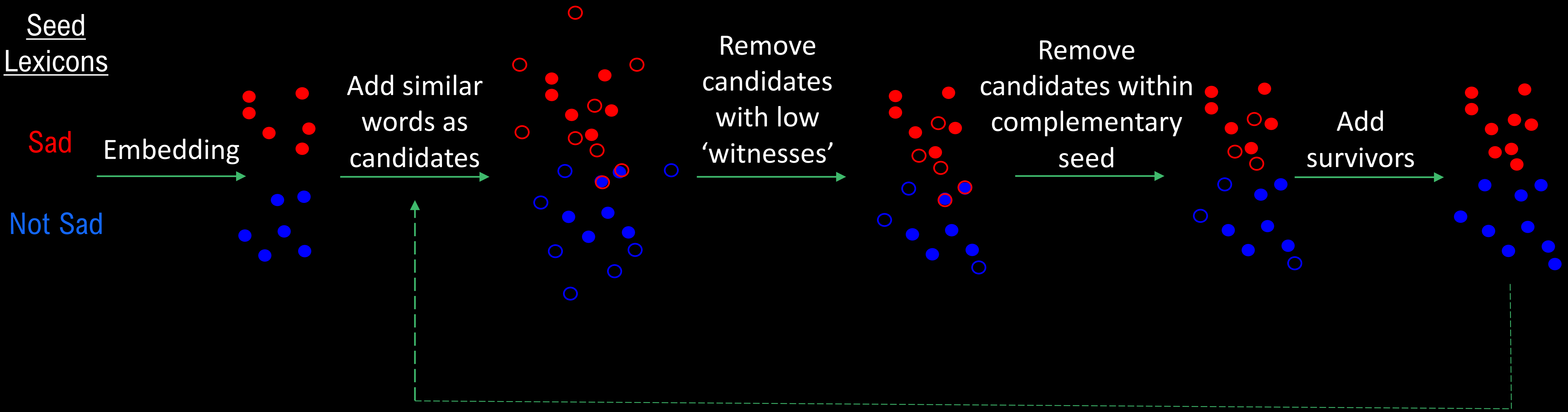
Add similar words as candidates

...

(under construction)



Expansions



Seed Lexicons

Sad

Not Sad

Embedding

Add similar words as candidates

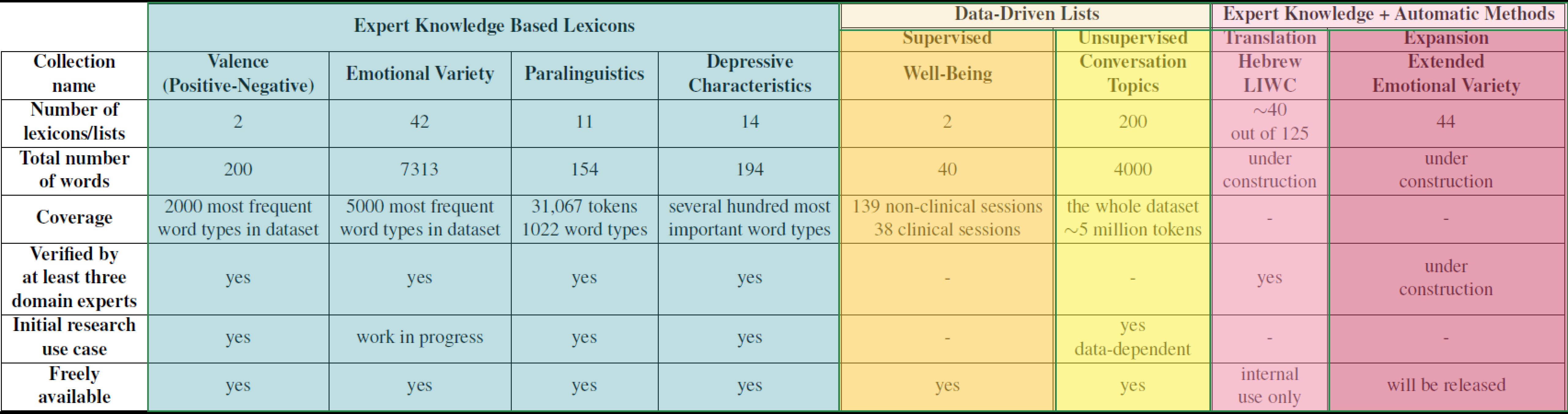
Remove candidates with low ‘witnesses’

Remove candidates within complementary seed

Add survivors

End

8 Lexicons Collections



	Expert Knowledge Based Lexicons				Data-Driven Lists		Expert Knowledge + Automatic Methods	
					Supervised	Unsupervised	Translation	Expansion
Collection name	**Valence (Positive-Negative)**	**Emotional Variety**	**Paralinguistics**	**Depressive Characteristics**	**Well-Being**	**Conversation Topics**	**Hebrew LIWC**	**Extended Emotional Variety**
Number of lexicons/lists	2	42	11	14	2	200	~40 out of 125	44
Total number of words	200	7313	154	194	40	4000	under construction	under construction
Coverage	2000 most frequent word types in dataset	5000 most frequent word types in dataset	31,067 tokens 1022 word types	several hundred most important word types	139 non-clinical sessions 38 clinical sessions	the whole dataset ~5 million tokens	-	-
Verified by at least three domain experts	yes	yes	yes	yes	-	-	yes	under construction
Initial research use case	yes	work in progress	yes	yes	-	yes data-dependent	-	-
Freely available	yes	yes	yes	yes	yes	yes	internal use only	will be released