

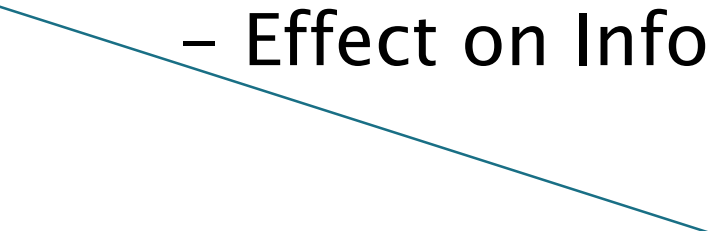
**Transliterated Pairs Acquisition in Medical Hebrew**  
or  
**Leveraging English resources to improve segmentation  
and POS tagging in Hebrew**

Click to edit Master subtitle style

Raphael Cohen, Yoav Goldberg and Michael Elhadad



# Outline

- Hebrew – what we need to know
  - Medical NLP– English Vs. Hebrew
  - Transliterations in the Medical Domain
  - Identifying Transliterations
  - Using English resources to improve Hebrew NLP:
    - Transliterated Pair Acquisition
    - Results
    - Effect on Segmentation, POS
    - Effect on Information Extraction
- 

# Introduction to Hebrew crash course\*

## ▶ Affixes

- ❖ *and, from, to, the, which, as, in* are prefixes
- ❖ *Possessives are suffixes to nouns*

**In her net → inernet**

## ➤ Unvocalized writing system:

- ❖ Vowels dropped when writing
- ❖ Diacritics are rarely written

**In her net → inernet → inhrnt**

## ➤ Rich morphology

- ❖ inernet could be further inflected into different forms based on sing/pl, masc/fem properties

\*Adapted from slides by Goldberg and Tsarfaty

# Introduction to Hebrew crash course – cont`

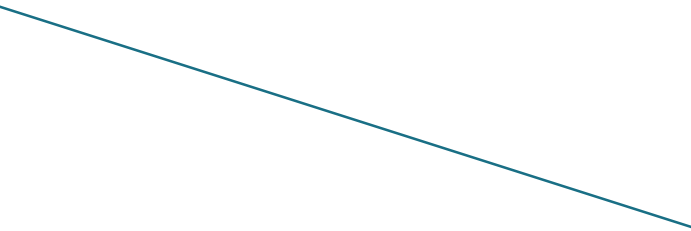
**To sum it up:**

- ▶ Complex, productive morphology
- ▶ Many word forms (487K distinct tokens in a 34M words corpus)
- ▶ High level of ambiguity
  - 2.7 tags/token, vs. 1.4 in English
- ▶ POS carries a lot of information:
  - Gender, number, tense, possessiveness...

# The Medical Domain

Adapted Resource	English	Hebrew
Lexicon	UMLS (800K concepts)	6,500 concepts
Ontology	UMLS (~1M Relations)	N/A
POS Tagger	98% [Tsuji Lab]	N/A
Parser	82.9% [Lease & Charniak]	N/A
Information Extraction	MetaMap, Medlee, cTAKE...	N/A

# What's the difference from Medical-English?

- ▶ Common Information Extraction approach: lexicon / ontology
  - ▶ In Hebrew:
    - transliterations
    - agglutination of suffixes and prefixes
    - “Smihut” form (construct state)
  - ▶ Difficult to extract terms and align concepts
- 

# Our Resources

“General” Hebrew (News Domain)

News  
Corpus

Dictionary  
(MILA)

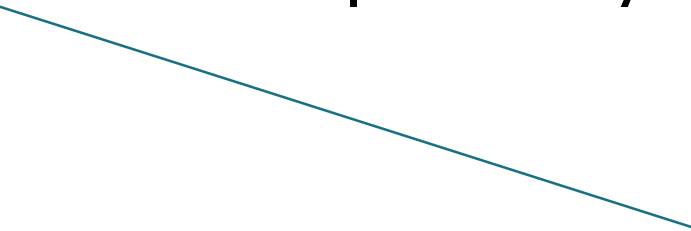
Morphological  
Disambiguator  
[Adler and  
Elhadad]

Medical Hebrew

The screenshot shows the WebMD website with the following elements:

- Header:** WebMD logo with the tagline "Better Information. Better Health." and a search bar.
- Navigation:** A horizontal menu with categories: TODAY'S NEWS, DISEASES & CONDITIONS, A-Z GUIDES, HEALTHY LIVING, HEALTH CARE SERVICES, PREGNANCY & FAMILY, and BOARDS & BLOGS.
- Main Content:**
  - TV Doctor Is In:** A featured article with a photo of actor Hugh Laurie and the text: "Is actor Hugh Laurie anything like his cranky on-screen alter ego?"
  - celebrity health:** A sub-section header.
  - Weight Loss:** A "WebMD Weight Loss CLINIC" advertisement for a personalized weight loss program.
  - SYMPTOM checker:** A tool for users to input symptoms, with options for MALE and FEMALE.
  - FRID IT FAST:** A list of health news items: "Huffer's High Hits Brain Hard", "Parents' Flanking Discipline", and "Drugs Slow Kidney Cancer".
  - Now Playing:** A section for "Men's Skin Care: Guys, focus on the essentials."
  - WATCH VIDEO:** A video player area.
  - TOP 12 Health Topics:** A list including "Anxiety".
  - Latest Headlines:** A list of news items: "House OK's Medicare Drug Negotiations", "Trans Fats May Increase Infertility", and "Not Overweight? You May Still Be Fat".
  - Counting Candles:** An advertisement for a quiz: "How old will you live to be?"

# Unknown words and domain adaptation

- ▶ The medical domain in English presents many unknown tokens
  - ▶ Algorithms designed to adapt to the medical domain mention unknown words to be the major source of error [Lease and Charniak, Blitzer]
  - ▶ The Morphological Disambiguator is based on MILA's Morphological Analyzer which is hampered by unknown words
- 



# Unknowns and our corpora

	Infomed	Neurology
Token Types	74,030	12,896
Unknown	16,514	3,919
% of unknown	~20%	~30%

Many of the unknowns are transliterated (Later)



# What are transliterations?

- ▶ Translation by sound: translating a word phonetically, according to its sound instead of its meaning
- ▶ Common names, brand names, but sometimes not proper and



# When are words transliterated?

- ▶ Names are usually transliterated.
- ▶ Technical words missing in the target language.
- ▶ In medical-Hebrew, many words are transliterated in textbooks
- ▶ Just to sound like Hebrew  
-> o-'g-m-t)



# Transliterations and our Corpora

## Distribution of Unknown Tokens

▶ Info

,ים

רר

▶ Neu

וצ

# Transliterations and Agglutination

- ▶ Transliterated words acquire agglutinations (suffixes and prefixes) like regular Hebrew words.
- ▶ For example:
  - “Stent” → “סטנט”
  - “Stents” → “סטנט-ים” (the Hebrew plural suffix is acquired).
  - “The stent” → “ה-סטנט” (“The” is translated as the prefix “ה”).

**How do transliterations affect us?**



# Segmentation Faults

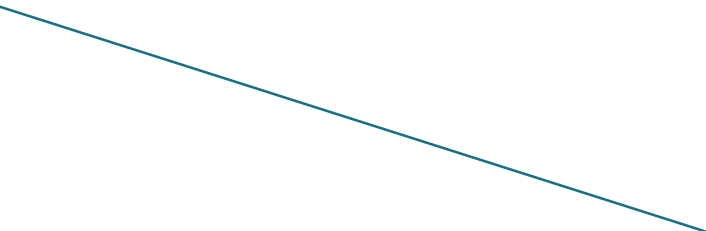
Token	Suggested Segmentation	Correct Segmentation
-------	------------------------	----------------------

# Part of Speech errors

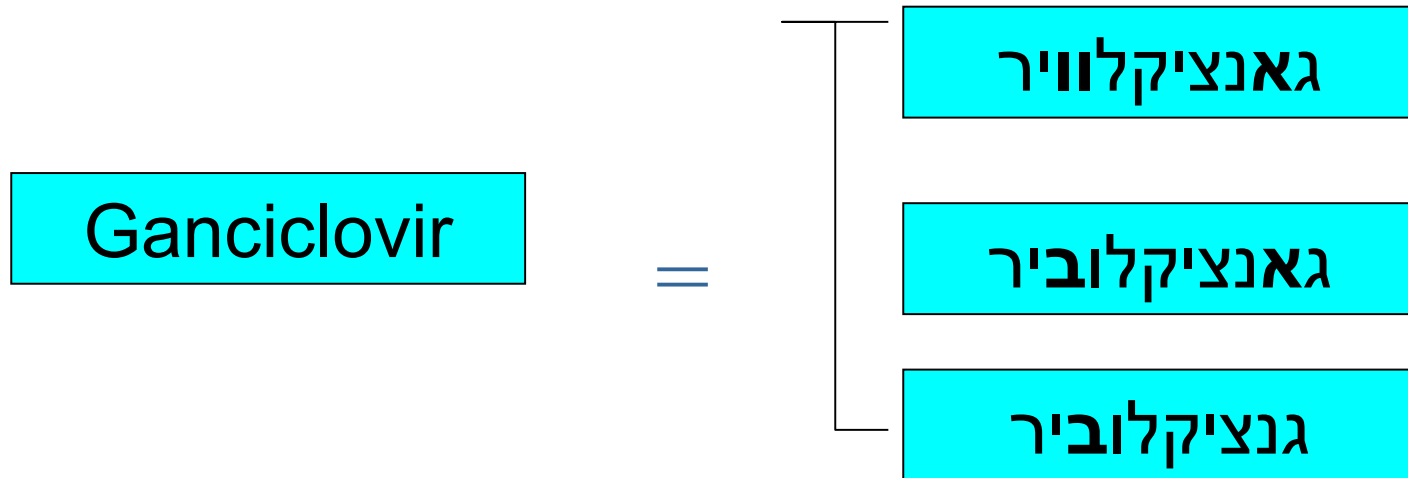
Token	POS Tag
המיפלגיה (hemiplegia)	
ספסטית (spastic)	
עם (with)	:PREPOSITION:
קונטרקטורות (contractures)	
בעיקר (mostly)	:ADVERB:
ביד (in the right)	PREPOSITION:NOUN-F,S,CONST:
ימין (hand)	:NOUN-M,S,ABS:



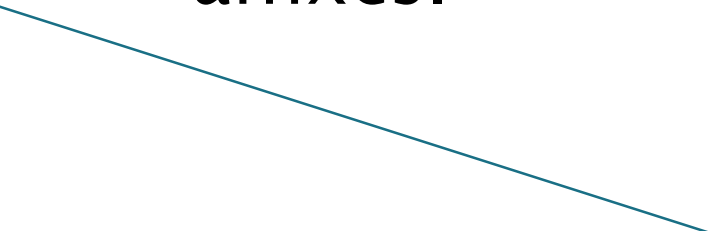
# To sum it up

- ▶ Transliterations are abundant in Medical-Hebrew
  - ▶ Transliterations confuse NLP tools.
  - ▶ Most segmentation errors (~90%) are due to transliterated words.
  - ▶ Let's get to work...
- 

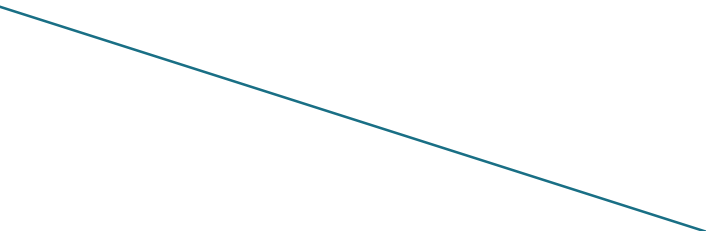
# Transliterated pair acquisition



# Related Work

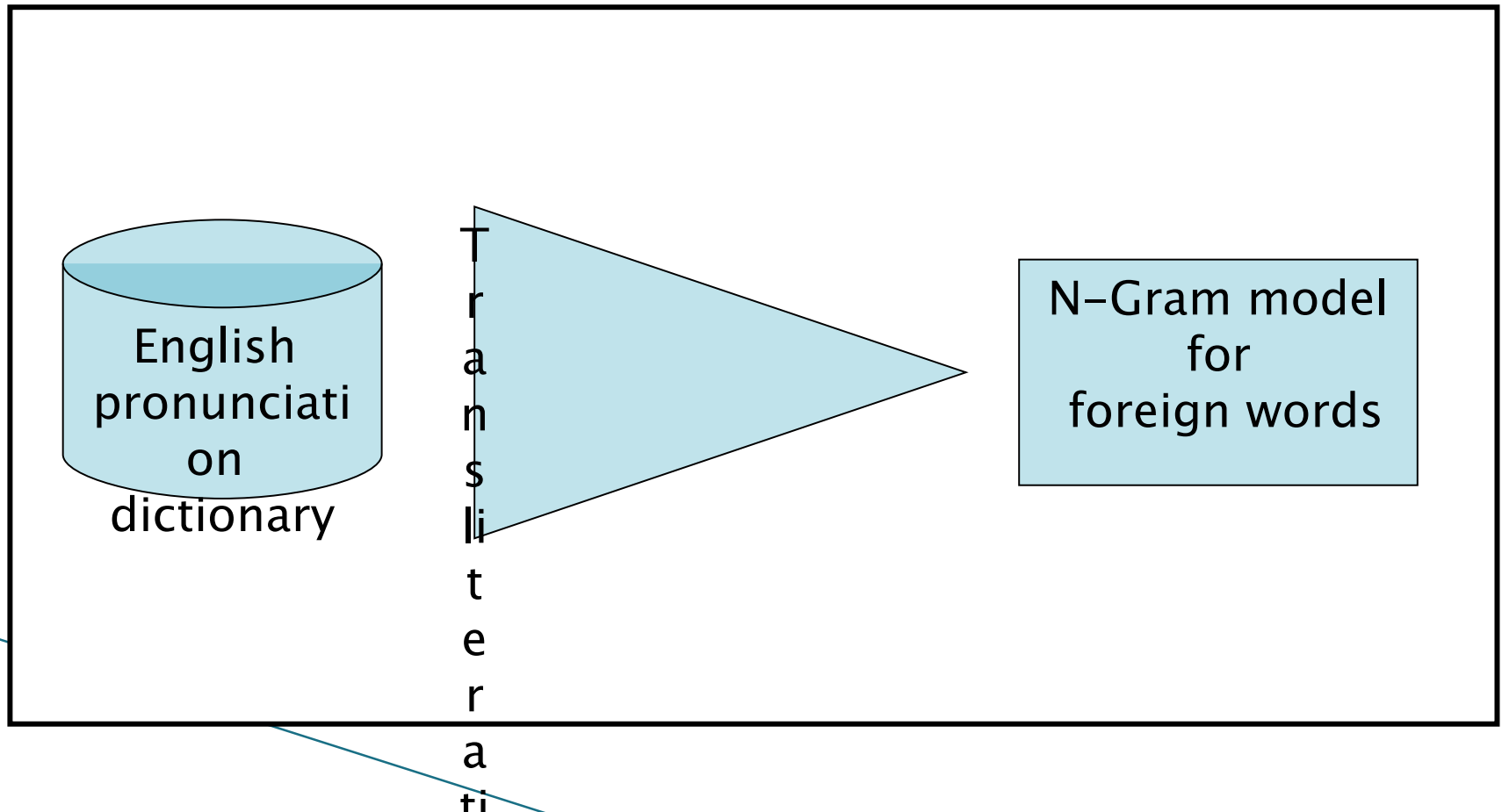
- ▶ Kirschenbaum and Wintner (2009) – transliteration from Hebrew to English for machine translation.
  - ▶ Kirschenbaum and Wintner (2010) – Pair acquisition from Wikipedia articles.
  - ▶ Goldberg and Elhadad (2008) – Transliteration identification. Not treating affixes.
- 

# Transliteration identification

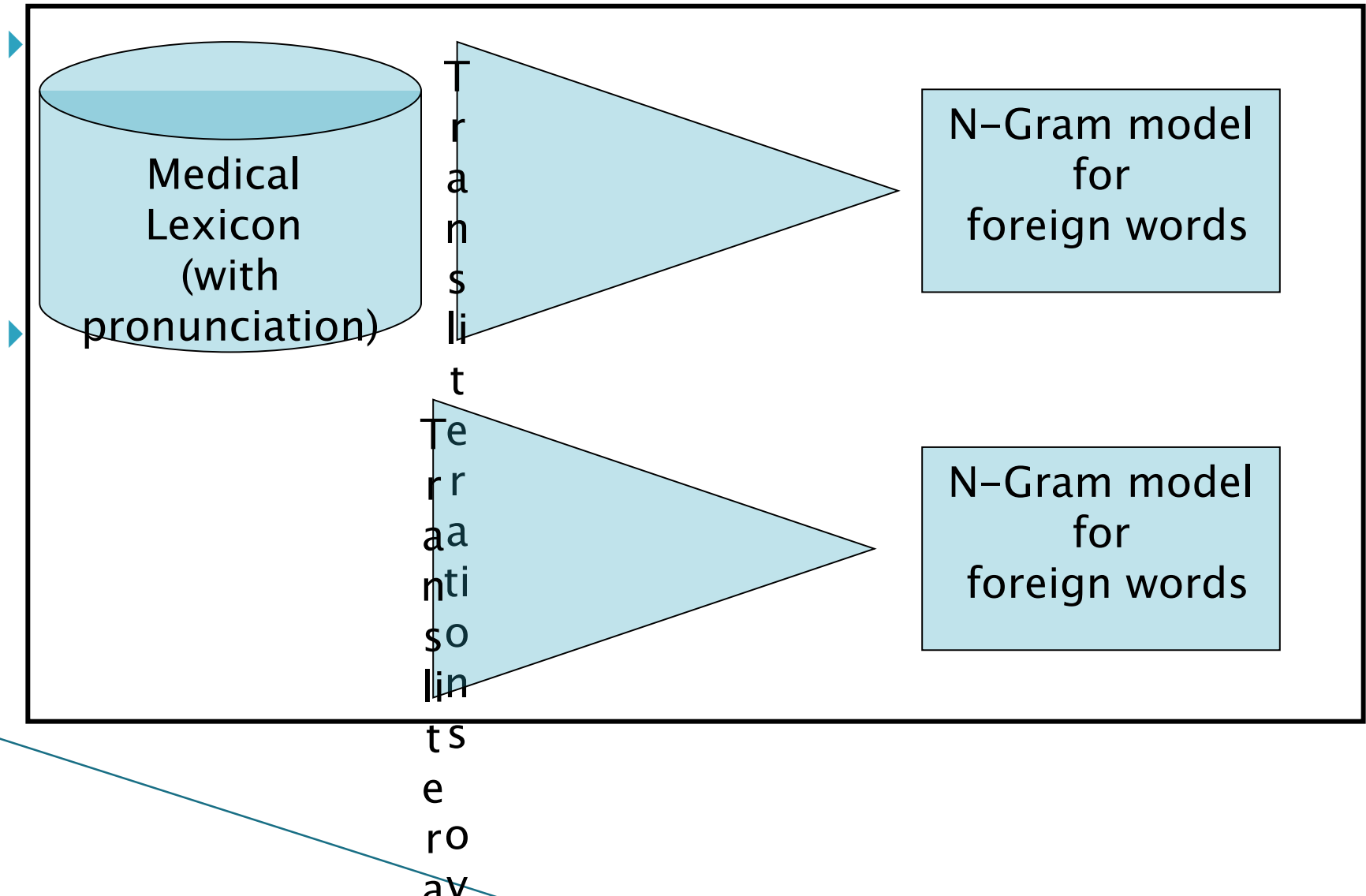
- Our practical definition:  
Words originating in Western languages which are not present in our lexicon.
  - “Strategy” = “אסטרטגיה” is already in the lexicon.
- 

# Identifying transliterations

- We extended the semi supervised method of Goldberg et al 2008



# Adaptation to the Medical Domain



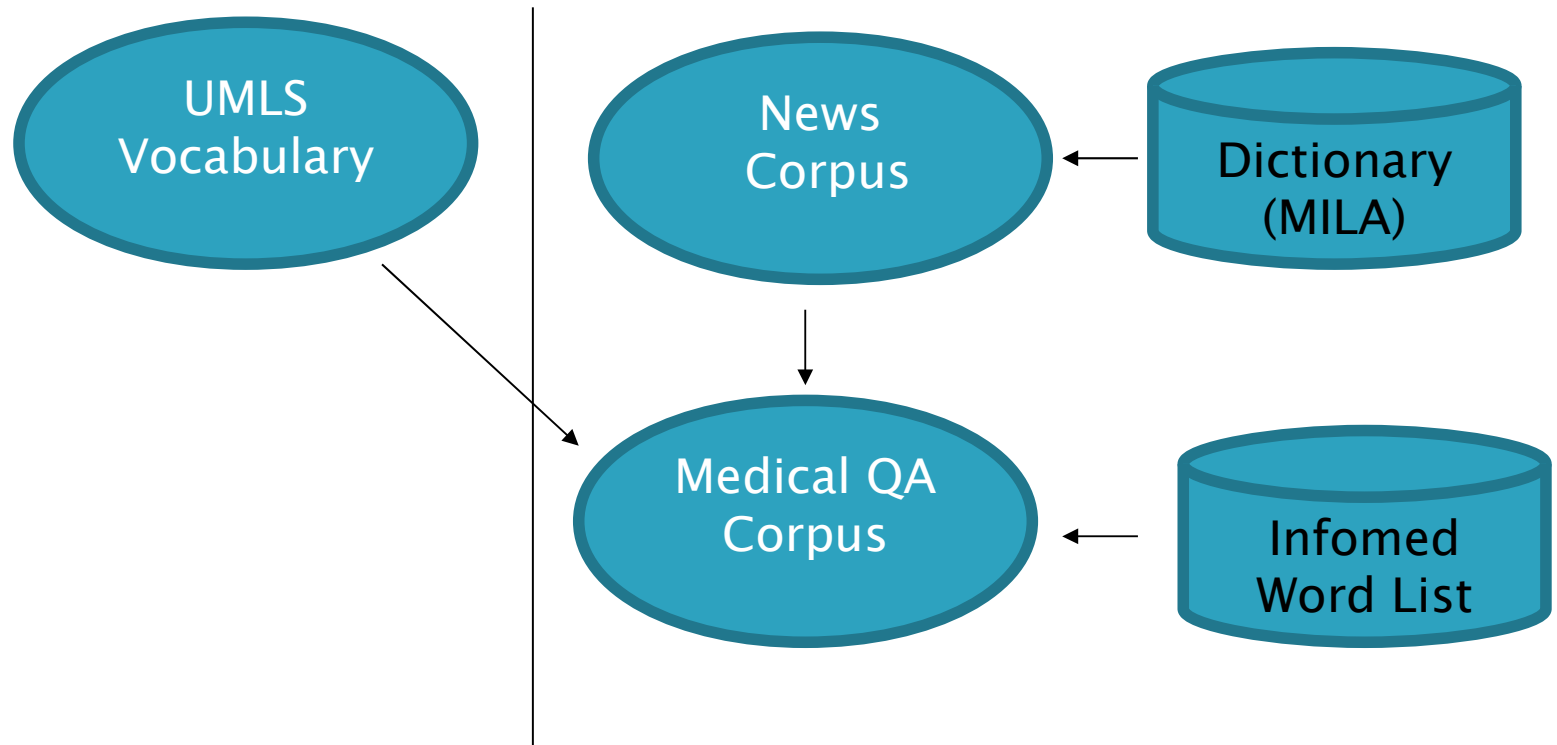
# Results (F-Score)

## Identifying transliterations

Medical

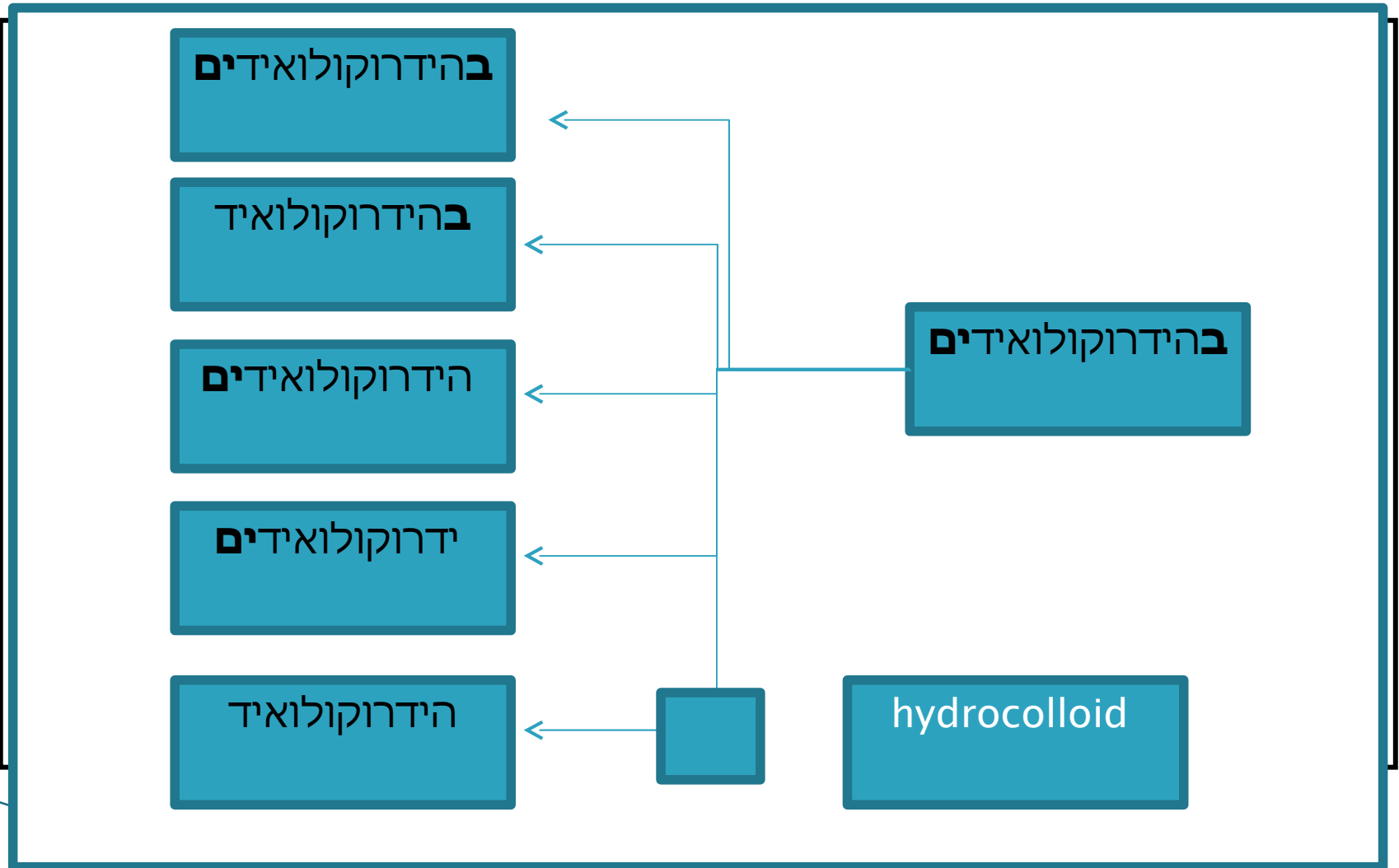


# Leveraging English resources to improve Hebrew NLP

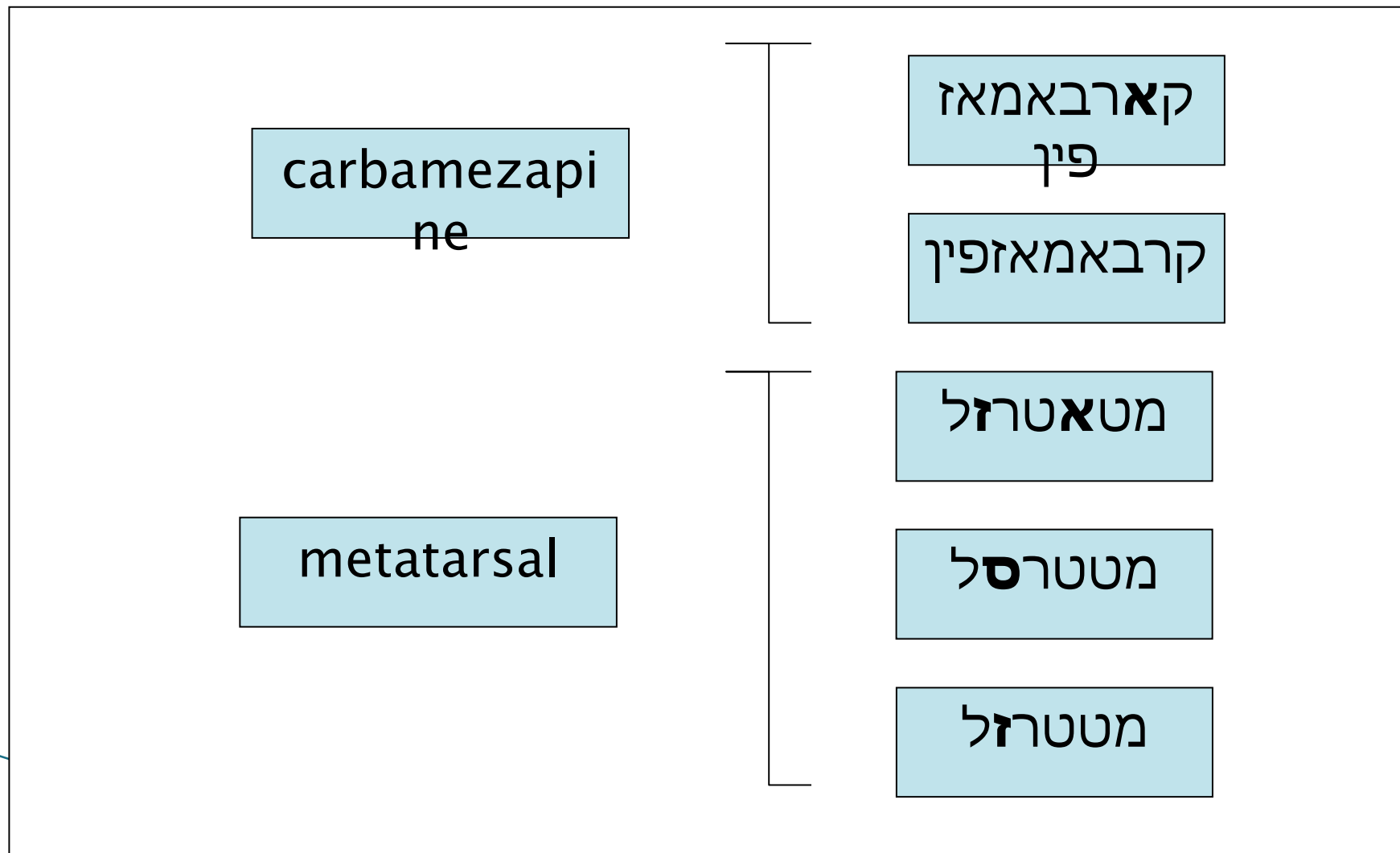




# Transliterated Pair Acquisition



# Results



# Effect on Segmentation / POS

	Infomed (Medical QA)	Neurology
Segmentation	49% error reduction (41 / 83)	44% error reduction (28 / 63)
POS	7.6% error reduction (15 / 196), 30% on foreign	16% error reduction (24 / 151)

# Results – Neurology domain

- ▶ 3,896 new transliterated pairs suggested
- ▶ 1,993 of these were new
- ▶ 30% of the correct pairs appeared only with an affix
- ▶ 336 added to the medical Lexicon

Effect on:	Neurology
Segmentation	51% total error reduction (33/64)
POS	20% total error reduction (31/151)

# Impact on Morphological Analyzer

Token	POS Tag
<b>המיפלגיה</b> (hemiplegia)	
<b>ספסטית</b> (spastic)	
<b>עם</b> (with)	:PREPOSITION:
<b>קונטרקטורות</b> (contractures)	
<b>בעיקר</b> (mostly)	:ADVERB:
<b>ביד</b> (in the right)	PREPOSITION:NOUN-F,S,CONST:
<b>ימין</b> (hand)	:NOUN-M,S,ABS:

# Impact on Morphological Analyzer

Token	POS Tag
המיפליגיה (hemiplegia)	DEF:PARTICIPLE-M,P,A,ABS: ✓
ספסטית (spastic)	:NOUN-F,S,ABS: ✓
עם (with)	:PREPOSITION:
קונטרקטורות (contractures)	:NOUN-M,S,ABS: ✓
בעיקר (mostly)	:ADVERB:
ביד (in the right)	PREPOSITION:NOUN-F,S,CONST:
ימין (hand)	:NOUN-M,S,ABS:

# Impact on Information Extraction

Infomed (Medical QA) – annotated corpus:

Recall                    56%    -> 60.4%

Precision                86.1% -> 87%

F-Score                  68%    -> 71.3%

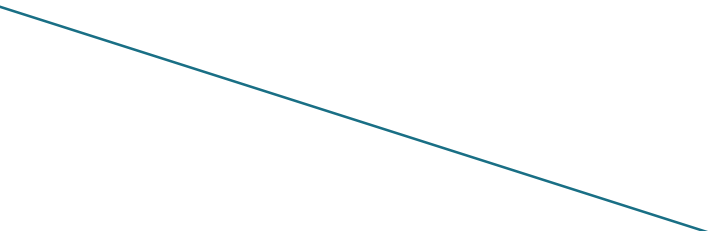
Neurology – not annotated:

20% increase in number of identified concepts.

7,746 -> 9,181 concepts

# Conclusions

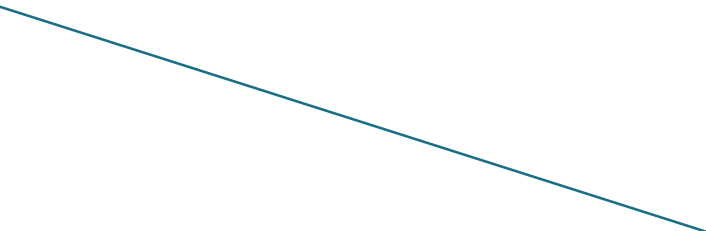
## Domain Adaptation of Morphological–Disambiguator

- ▶ News Domain → Medical Domain
  - ▶ Identified unknown words as major issue for domain adaptation
  - ▶ Identified transliteration as major source of unknown word types
- 



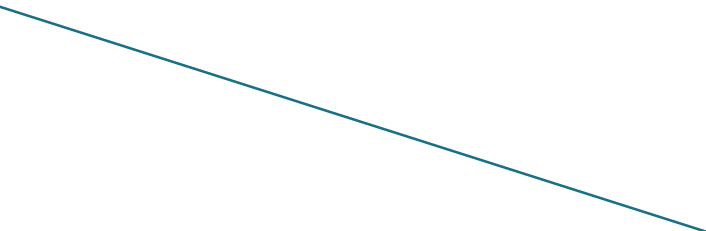
# Conclusions

We presented a method to acquire transliterated pairs in medical domain

- ▶ UMLS data source
  - ▶ Non-annotated Hebrew text (Infomed, Patient Records)
  - ▶ Combined to acquire transliterated pairs
- 

# Conclusions

Transliterated pairs help us

- ▶ Reduce POS/Morphological analysis (error reduction ~50%)
  - ▶ Improve information extraction (increased recall)
- 

**Thank You**



# LDA

- 2 שְׁתֵּן דָּרָךְ עֲרֵמוֹנִית מִתֵּן שֵׁלֶפֶחַיִת כְּלִיָּה אֹרֻלוֹג צְרִיכָה הַגְּדֵלָה תִּרְבִּית תְּכִיפֹת דְּחִיפָה שׁוֹפְכָה גָּבֵר אִיסוֹף תְּדִירוֹת בְּרִיחָה הַשְּׁתַנָּה הַטְּלָה פִּין
- 3 מִחֲזוֹר בִּיּוֹץ הַרְיֹן נֶסֶת שַׁחֲלָה אַחֹר זָקִיק מוֹעֵד יוֹם פּוֹרִיּוֹת סִדִּירוֹת עֲרָכָה בִּיצִית תֶּאֱרִיף קִבְלָה נֶסֶת כְּנִיסָה הַפְּרִיָה גִּינְקוֹלוֹג עֲבִי
- 4 וִיטְמִין בְּרוֹזֶל תְּמַצָּה חֶסֶר חֶסֶר רְמָה סְפִיגָה תוֹסֵף מְחֻסָּר סִיד תְּזוֹנָה דָּם גּוֹף מְזוֹן עוֹדֵף מִינְרָל מְאָגֵר אֲבָץ תוֹסֶפֶת כְּמוֹת
- 5 עֵבֶר סִפְנָה רוּחַ כֵּן אֲבַעְבוּעוֹת תְּשֻׁשׁ הוֹפְעָה זְמַן תַּחֲלָה בְּחוֹר עַד אִישׁ שְׁאֵלָה שֶׁלִּבְקַת מְנָה רָגַע עֲבָדָה שְׁנֵי סְבִירוֹת תַּיִם
- 6 לֵב עוֹרֵק מִסְתֵּם תָּזָה דָּם צְנִתוֹר חֶדֶר פָּגָם מְאָמֵץ טְרִשֶׁת אוֹטֵם קַרְדִּיּוֹלוֹג שְׁרִיר דָּפוֹ גוֹרֵם מַעֲקָף צְנִיחָה תְּמָרָה תְּמַצָּן כְּלִי  
טְרִשֶׁת בְּטְרִשֶׁת לְטְרִשֶׁת וְטְרִשֶׁת טְרִשֶׁתִי הַטְּרִשֶׁת
- 7 עֵין רוֹפֵא רְאִיָּה טְפָה מְקָרָה צְרָף דְּבָר טֶשֶׁטוּשׁ יִכָּשׁ חֶלֶק רֶשֶׁתִית מְסַפֵּר דְּמַעָה עֲפָעָף פִּגִיעָה עֲדָשָׁה פְּנִים אוֹר סִימְפֻטוֹם נֹזֵק
- 8 רְאָה תָּזָה נְשִׁימָה צִילוֹם שְׁעוֹל קֶצֶר אוֹיֵר אֶסְתֵּמָה שְׁחֶפֶת לַחָה צִלַּע בֵּית תִּסְנִין סִמְפוֹן תְּמוֹנָה תְּמַצָּן צֶל בְּרוֹנְכִיטִיס הַצִּלְלָה נִפְחָ
- 9 מְצַב הַחֲמָרָה בַּד תוֹצְאָה תֵּאוֹר יַעוֹץ קְדָמוֹת סִבָּה תַּחֲלָה אֲרָף כֵּן פֶּתֶרוֹן הַדְּרָדְרוֹת אֲכַשְׁרוֹת מְשֻׁהוֹ טוֹבָה וְכִלְתָּ קָרוֹב עֲתִיד שְׁנֵי

# Impact on LDA (Neurology Corpus)

- ▶ “Febrile seizures” cluster:

אָרוֹעַ|חֹום|פּרָכּוּס (event|febrile|seizure)

- ▶ Contains “סטזוליד” (Stesolid), the treatment drug, only when run with the Transliterations Lexicon
- ▶ “סטזוליד” appears affixed 50% of the time and is not segmented correctly