

Agreement matters: Challenges of translating **into** a MRL

Click to edit Master slide background
Yoav Goldberg
Ben Gurion University

Why should we care about syntactic modeling of MRLs?

[Click to edit Master subtitle style](#)

A brief summary

A brief summary

- What Kevin said.

Example: English Hebrew



I wash the car

Google translate

Example: English Hebrew



I wash the car

Google translate

אני רוחץ את

(I wash the car)

Example: English Hebrew



I wash the car

Google translate

אני רוחץ את

(I wash the car)



I wash the floor

Google translate

Example: English [?] Hebrew



I wash the car

Google translate

אני רוחץ את

(I wash the car)



I wash the floor

Google translate

אני שוטפת את

(I wash the floor)

Example: English Hebrew



I wash the car

Google translate

אני רוחץ את

(I wash the car)



I wash the floor

Google translate

אני שוטפת את

(I wash the floor)



Example: English [?] Hebrew



I wash the car

Google

Masculine

אני רוחץ את

(I wash the car)



I wash the floor

Google

Feminine

אני שוטפת את

(I wash the floor)



Hebrew Fact 1

Hebrew Verbs are morphologically marked for Gender

(and Number, and Person, and Tense..)

Click to edit Master subtitle style

Example: English Hebrew



I wash the car

Google

Masculine

אני רוחץ את

(I wash the car)



I wash the floor

Google

Feminine

אני שוטפת את

(I wash the floor)



Are these bad translations?

Example: English [?] Hebrew



I washed

No. These are actually quite good.

Google

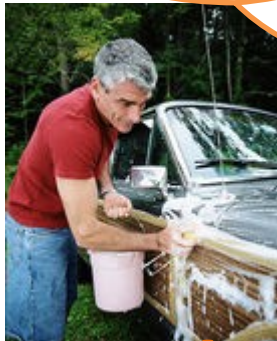
יאת

No gender information in source.

(I washed

Target must indicate gender.

[?] translator uses world knowledge.



Are these bad translations?

Let's have some fun

Click to edit Master subtitle style

Language Models as Social Indicators

- I love her
- I love him




Language Models as Social Indicators

- I love her
- I love him


- אני אוהב אותה
- אני אוהבת אותו




Language Models as Social Indicators

- I love her
 - I love him
 - I love meat
 - I love vegetables
- 
- אני אוהב אותה
 - אני אוהבת אותו


Language Models as Social Indicators

- I love her
 - I love him
 - I love meat
 - I love vegetables
- 
- אני אוהב אותה
 - אני אוהבת אותו
 - אני אוהב בשר
 - אני אוהבת ירקות


Language Models as Social Indicators

- I love her
 - I love him
 - I love meat
 - I love vegetables
 - I love to eat
 - I love to cook
- 
- אני אוהב אותה
 - אני אוהבת אותו
 - אני אוהב בשר
 - אני אוהבת ירקות


Language Models as Social Indicators

- I love her
 - I love him
 - I love meat
 - I love vegetables
 - I love to eat
 - I love to cook
- 
- אני אוהב אותה
 - אני אוהבת אותו
 - אני אוהב בשר
 - אני אוהבת ירקות
 - אני אוהב לאכול
 - אני אוהבת לבשל

Language Models as Social Indicators

- I love her
 - I love him
 - I love meat
 - I love vegetables
 - I love to eat
 - I love to cook
 - I love hash
 - I love marijuana
- 
- אני אוהב אותה
 - אני אוהבת אותו
 - אני אוהב בשר
 - אני אוהבת ירקות
 - אני אוהב לאכול
 - אני אוהבת לבשל

Language Models as Social Indicators

- I love her
 - I love him
 - I love meat
 - I love vegetables 
 - I love to eat
 - I love to cook
 - I love hash
 - I love marijuana
- אני אוהב אותה
 - אני אוהבת אותו
 - אני אוהב בשר
 - אני אוהב ירקות
 - אני אוהב לאכול
 - אני אוהבת לבשל
 - אני אוהב חשיש
 - אני אוהבת מריחואנה

Language Models as Social Indicators

- I hate him.



- אני שונאת אותו.

Language Models as Social Indicators

- I hate him.
- I hate her.



- אני שונאת אותו.

Language Models as Social Indicators

- I hate him.
- I hate her.



- אני שונאת אותו.
- אני שונאת אותה.

Language Models as Social Indicators

- I hate him.
- I hate her.
- I hate him



- אני שונאת אותו.
- אני שונאת אותה.
- אני שונא אותו.

Language Models as Social Indicators

- I hate him.
- I hate her.
- I hate him



- אני שונאת אותו.
- אני שונאת אותה.
- אני שונא אותו.

Really? A dot?!
Not very stable...

Language Models as Social Indicators


- I hate him.
- I hate her.
- I hate him
- I hate her



- אני שונאת אותו.
- אני שונאת אותה.
- אני שונא אותו.
- אני שונאת אותה.

Really? A dot?!
Not very stable...

Language Models as Social Indicators

- I hate him.
 - I hate her.
 - I hate him
 - I hate her
- 
- אני שונאת אותו.
 - אני שונאת אותה.
 - אני שונא אותו
 - אני שונאת אותה

Hmm... is there
a message
here after all?

Really? A dot?!
Not very stable...

Language Models as Social Indicators

- I love
- I hate



Language Models as Social Indicators

- I love
- I hate

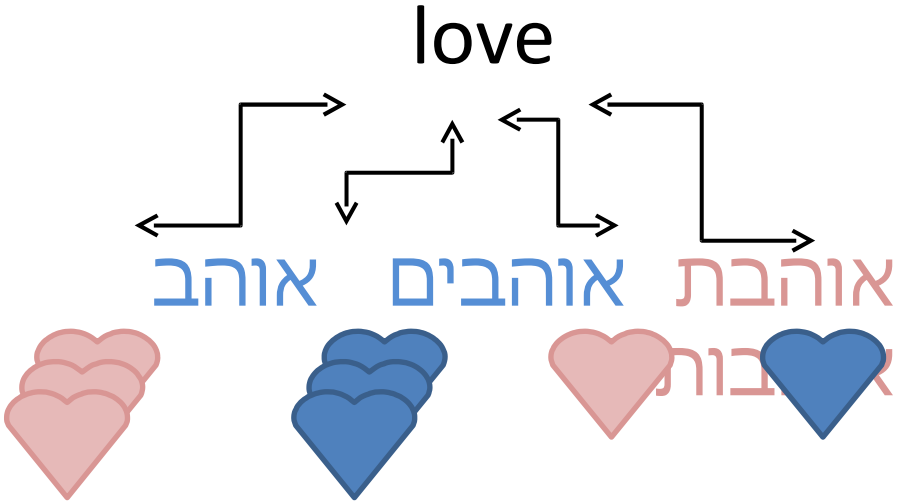


- אני אוהב
- אני שונאת

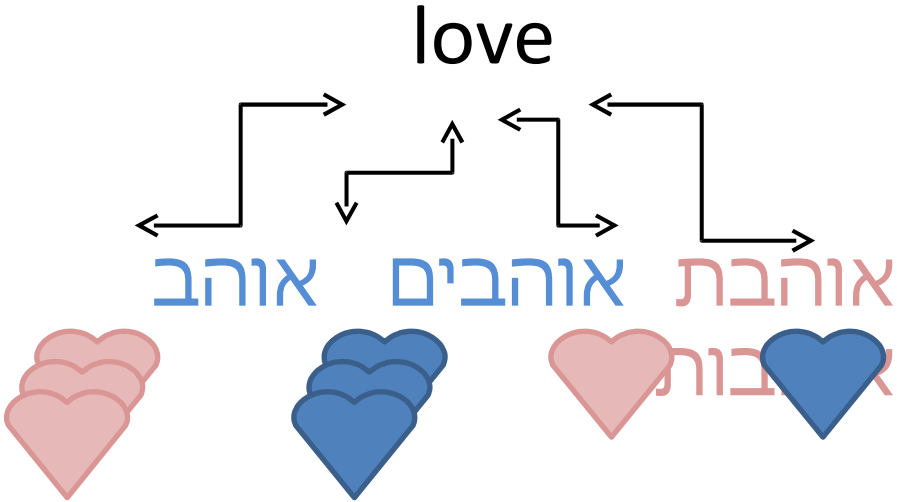
Back to Machine Translation

Click to edit Master subtitle style

One English [??] Many Hebrew

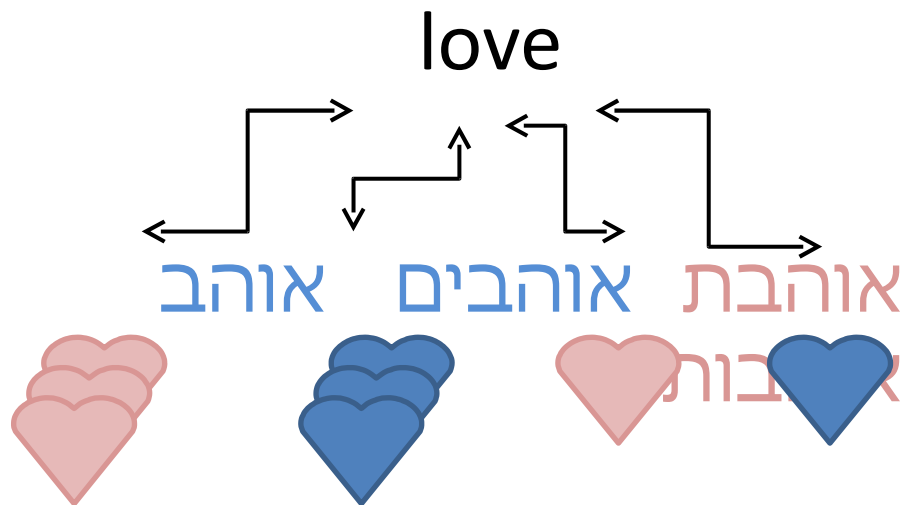


One English [??] Many Hebrew



Need to acquire more knowledge

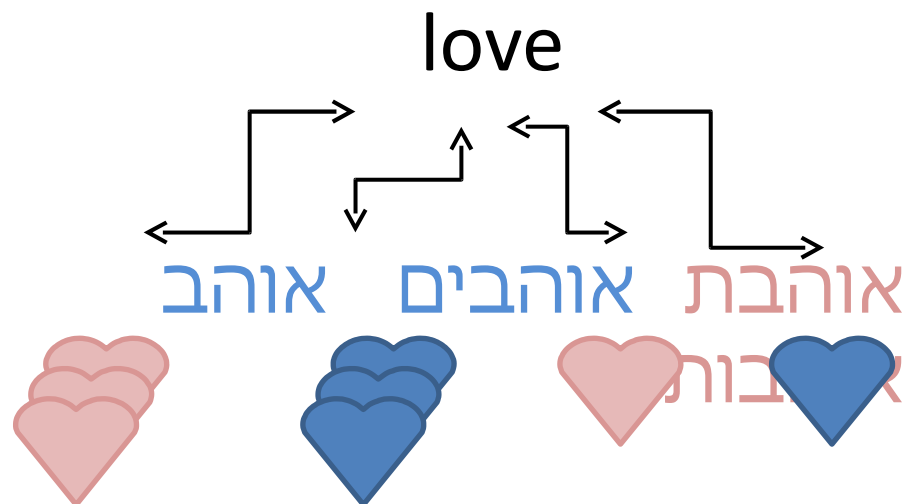
One English [??] Many Hebrew



Need to acquire more knowledge

- ... use larger parallel corpora

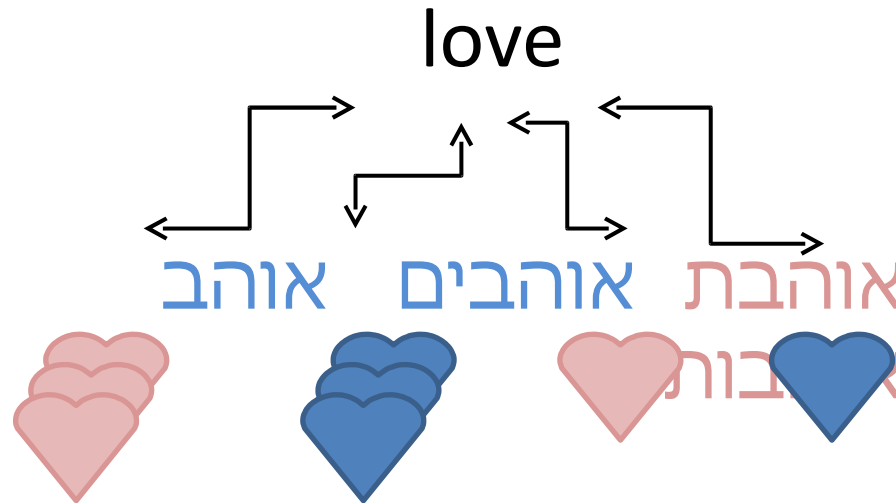
One English [??] Many Hebrew



Need to acquire more knowledge

- ... use larger parallel corpora
- use dictionaries

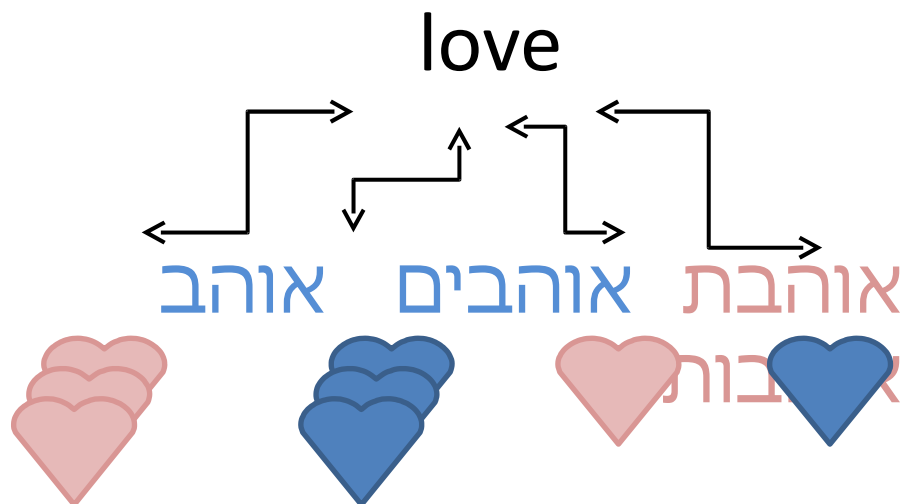
One English [??] Many Hebrew



Need to acquire more knowledge

- ... use larger parallel corpora
- ... use dictionaries
- ... use FSA to model inflections

One English [??] Many Hebrew



Need to acquire more knowledge

- ... use larger parallel corpora
- ... use dictionaries
- ... use FSAs to model inflections
- **Let's assume this is solved**

One English [?][?] Many Hebrew

Hebrew [?] English:

easy

אוהבת אני □

love

אוהב אני □

One English [?][?] Many Hebrew

Hebrew [?] English:

easy

אוהבת אני □

love

love □ אוהב ערו

One English [?][?] Many Hebrew

Hebrew [?] English:

easy

אוהבת אני [?]

love

love [?] אוהבת אני

Don't worry about it.

Just say "love".

The reader will

decide.

One English [?][?] Many Hebrew

Hebrew [?] English:
easy

אני אוהבת אני □

love

אני אוהב אני □ love

English [?] Hebrew: hard
אני אוהבים אני □ We

love I love [?][?] אני אוהב

אני אוהבת [?][?] אני אוהבות אני □ We
אני אוהבים [?][?] אני אוהבות [?][?] We

love

אני אוהבות [?][?]

Don't worry about it.
Just say "love".
The reader will
decide.

One English [?][?] Many Hebrew

Hebrew [?] English:
easy

אוהבת אני □

love

אוהב אני □ love

English [?] Hebrew: hard
אוהבים אני □ We

love I love [?][?] אני אוהב [?][?]

אוהבת אני [?][?] אוהבות אני [?][?]
אוהבים אני [?][?] אוהבות אני [?][?]

love

אני אוהבות [?][?]

Don't worry about it.
Just say "love".
The reader will
decide.

Which form to choose?
Translator must decide.

HOW??

When translating into an MRL:

- Many possible word forms
 - Hard to acquire [but assume its solved]
- Need to choose correct inflection

One English [?] Many Hebrew

Hebrew [?] English:
easy

אוהבת אני □

love

אוהב אני □ love

English [?] Hebrew: hard
אוהבים אני □ We

love I love [?] אוהב אני

אוהבת אני [?] אוהבות אני □ We
אוהבים אני [?] אוהבים אני

love

אוהבות אני [?]

Don't worry about it.
Just say "love".
The reader will
decide.

Which form to choose?
Translator must decide.

HOW??

One English [?] Many Hebrew

Hebrew [?] English:

easy

אהבת אני

love

אהב אני

English [?] Hebrew: hard
אהב אני

love | אהב אני [?] | love I

אהבת אני [?] | אהבות אני
אהב אני [?] | אהבים אני

love | אהבות אני [?] | love

Don't worry about it.
Just say "love".

Just choose one at random?

In the worst case we'll insult
someone..

Which form to choose?

Translator must decide.

HOW??

Hebrew Fact 2

Hebrew Verbs **agree** with Subject on gender and number

[Click to edit Master subtitle style](#)

Agreement dictates form

אני אוהב ? I love

אני אוהבת ?

אני אוהבים ?

אני אוהבות ?

Agreement dictates form

אני אוהב ?
I love ?
singular singular

אני אוהבת ?
singular

אני אוהבים ?
singular

אני אוהבות ?
singular

Agreement dictates form


I love אני אוהב
singular singular singular

אני אוהבת
singular singular

אני אוהבים
plural singular

אני אוהבות
plural singular

Agreement dictates form

I love אני אוהב
singular singular singular 

אני אוהבת
singular singular 

אני אוהבים
plural singular 

אני אוהבות
plural singular 

no missing information

The girls love אומץ הבחורות אומץ
plur fem

הבחורות אוהבת אומץ ?

הבחורות אוהבים אומץ ?

הבחורות אוהבות אומץ ?

no missing information

The girls love הבחורות ?
plural / fem sing / masc plural / fem

הבחורות אוהבת ?
plural / fem sing / fem

הבחורות אוהבים ?
plural / fem plural / masc

הבחורות אוהבות ?
plural / fem plural / fem

no missing information

The girls love
אומרים

plural / fem

הבחורות ??

sing / masc

plural / fem



הבחורות אוהבת ??

sing / fem

plural / fem



הבחורות אוהבים ??

plural / masc

plural / fem



הבחורות אוהבות ??

plural / fem

plural / fem



When translating into an MRL:

- Many possible word forms
 - Hard to acquire [but assume its solved]
- Need to choose correct inflection
- Inflection is determined based on information which is **external** to the word

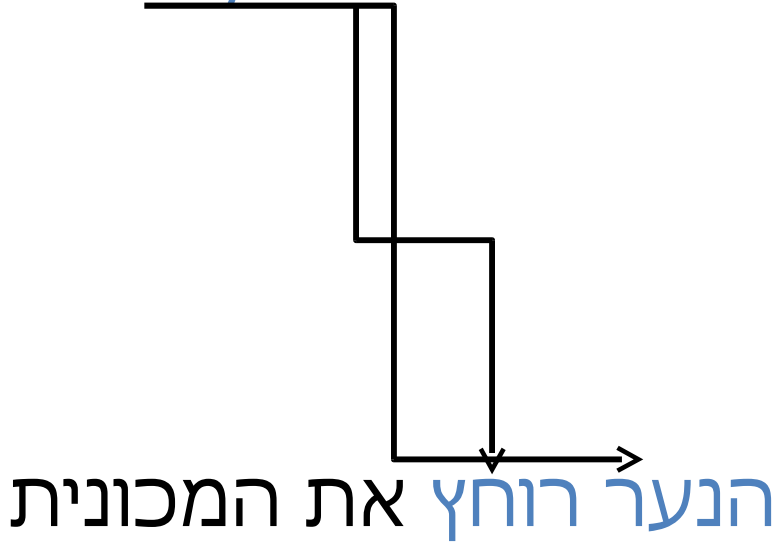
Back to Google Translate

The boy washes the car

The girl washes the car

Back to Google Translate

The boy washes the car

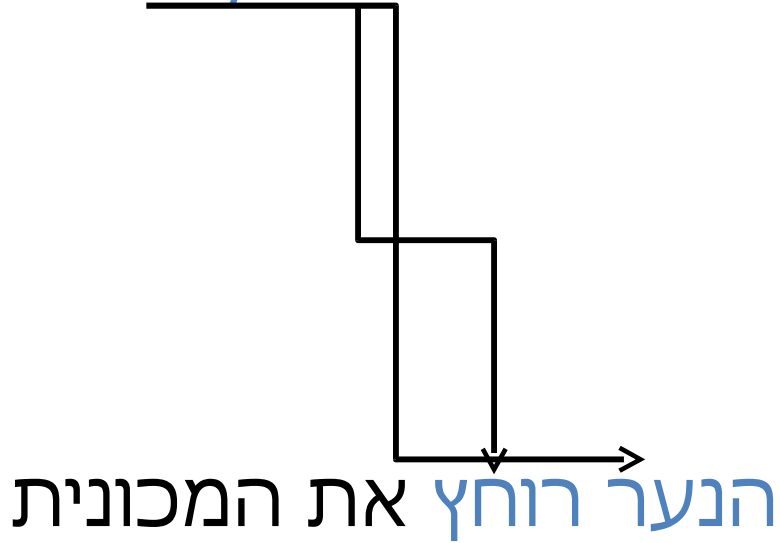


The girl washes the car



Back to Google Translate

The boy washes the car

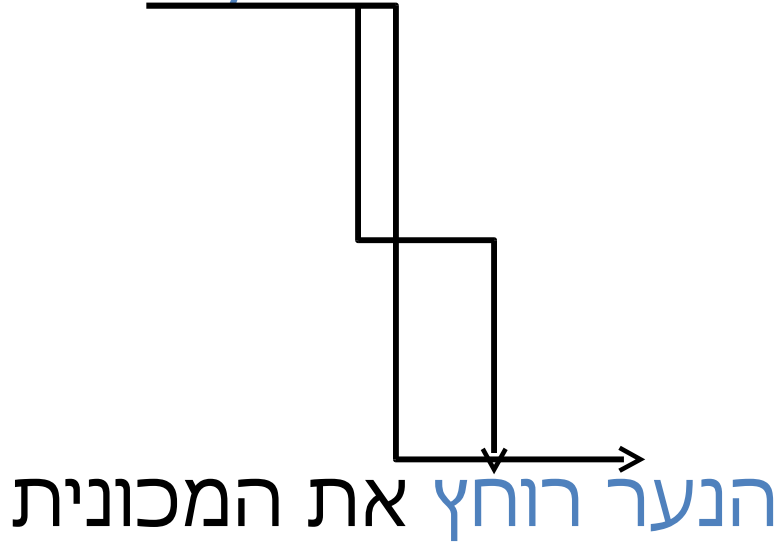


The girl washes the car



Back to Google Translate

The boy washes the car



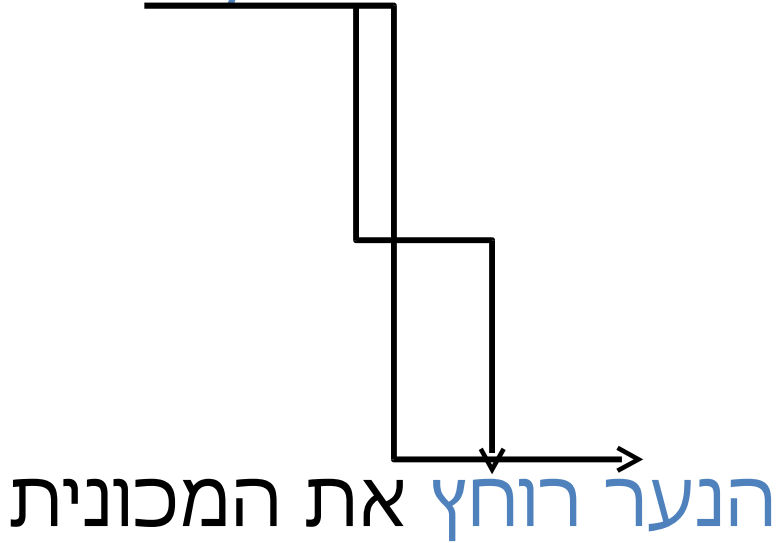
The girl washes the car



Good job Franz!

Back to Google Translate

The boy washes the car



The girl washes the car



Good job Franz?

Back to Google Translate

The boy washes the car



The girl washes the car



Good job Franz?

Back to Google Translate

The boy with the sunglasses washes the floor

הנער עם משקפי השמש שוטפת את הרצפה

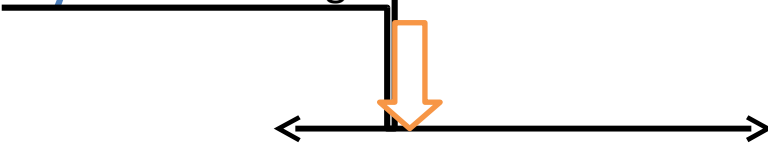
The girl with the sunglasses washes the car

:

.....
Good job Franz?

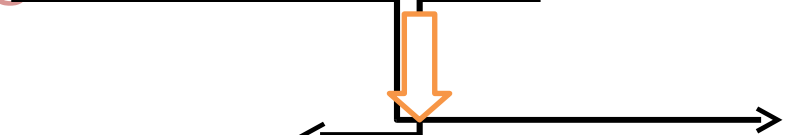
Back to Google Translate

The boy with the sunglasses washes the floor



הנער עם משקפי השמש שוטפת את הרצפה

The girl with the sunglasses washes the car

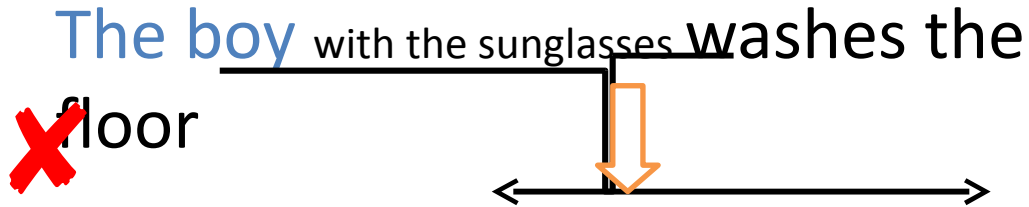


הבחורה עם משקפי השמש שוטף את המכונית

Good job Franz?

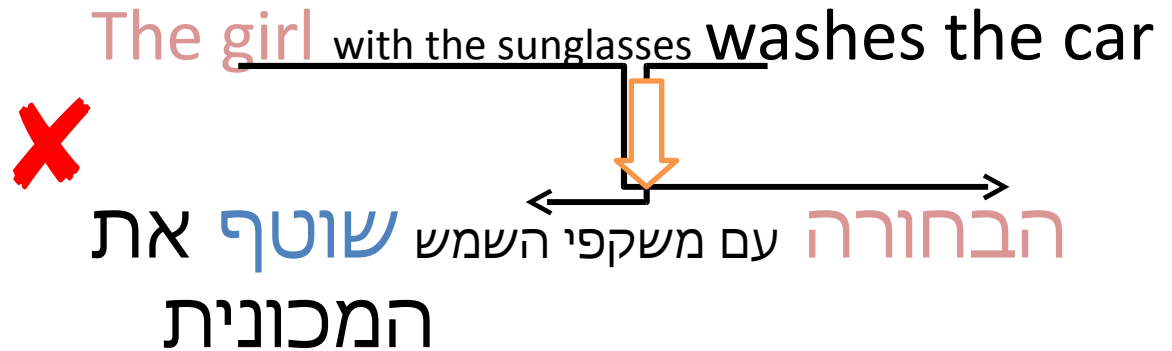
Back to Google Translate

The boy with the sunglasses washes the
floor



הנער עם משקפי השמש שוטפת את
הרצפה

The girl with the sunglasses washes the car

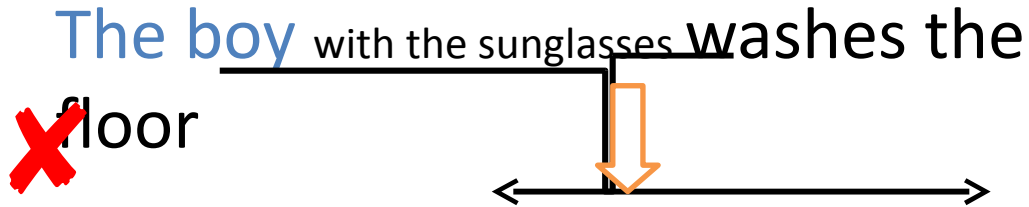


הבחורה עם משקפי השמש שוטף את
המכונית

Good job Franz?

Back to Google Translate

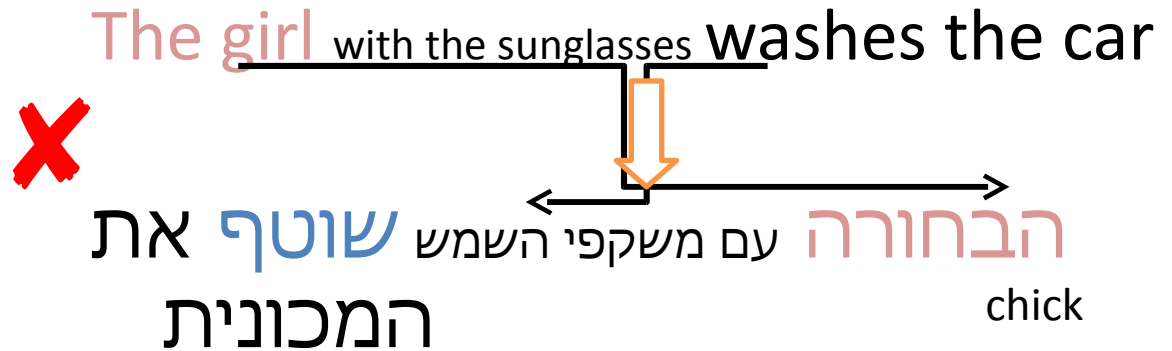
The boy with the sunglasses washes the
floor



הנער עם משקפי השמש שוטפת את
הרצפה

young-man

The girl with the sunglasses washes the car



הבחורה עם משקפי השמש שוטף את
המכונית

chick

Good job Franz?

What happened?

- Long distance agreement
- Can't be represented in phrase-table
- Can't be represented in n-gram LM
- Local “semantic” information from LM/Phrase
- Bad translation (ungrammatical)

What happened?

- Long distance agreement
- Can't be represented in phrase-table
- Can't be represented in n-gram LM
- Local “semantic” information from LM/Phrase
- Bad translation (ungrammatical)

It's not Franz's fault, but the system's

When translating into an MRL:

- Many possible word forms
 - Hard to acquire [but I assume its solved]
- Need to choose correct inflection
- Inflection is determined based on information which is **external** to the word **and frequently far away from it**

S-V Dep-Length	Count	Percent
1	3218	42%
2	1504	19%
3	914	12%
4	405	5%
5	297	4%
>5	1322	17%

Distance from Verb to Subject in the Hebrew Dependency
Treebank (news domain)

which is **external** to the word **and frequently far away from it**

S-V Dep-Length	Count	Percent
1	3218	42%
2	1504	19%
3	914	12%
4	405	5%
5	297	4%
>5	1322	17%

Distance from Verb to Subject in the Hebrew Dependency
Treebank (news domain)

which is **external** to the word **and frequently far away from it**

2 words apart is already
though for reliably estimating
in an n-gram based system!

When translating into an MRL:

- Many possible word forms
 - Hard to acquire [but I assume its solved]
- Need to choose correct inflection
- Inflection is determined based on information which is **external** to the word **and frequently far away from it**

When translating into an MRL:

- Many possible word forms
 - Hard to acquire [but I assume its solved]
 - Need to choose correct inflection
 - Inflection is determined based on information which is **external** to the word **and frequently far away from it**
- ☐ Phrase based + N-gram LM can't do it**

What if both languages are MRLs?

- Gender/number marked on both sides
- No need for word-external information
- We can translate word \leftrightarrow word again
- MRL \leftrightarrow MRL is easy!

What if both languages are MRLs?

- Gender/number marked on both sides
- No need for word-external information
- We can translate word \leftrightarrow word again
- MRL \leftrightarrow MRL is easy!

Wrong!

What if both languages are MRLs?

- Gender/number marked on both sides
- **But:**
 - agreement patterns differ between languages
 - gender information differ between languages

What if both languages are MRLs?

Example:

- Spanish and Hebrew have **adjective-noun** agreement

What if both languages are MRLs?

Example:

- Spanish and Hebrew have **adjective-noun** agreement

– new shirt

- חולצה חדשה
- nueva [◊]camisa

What if both languages are MRLs?

Example:

- Spanish and Hebrew have **adjective-noun** agreement

– new shirt

- חולצה חדשה
- [◇]nueva camisa

– new car

- מכונית חדשה
- [◇]nuevo automovil

What if both languages are MRLs?

Example:

- Spanish and Hebrew have **adjective-noun** agreement

– new shirt

- חדשה חולצה
-  nueva camisa

– new computer

- חדש מחשב
-  nueva computadora

– new car

- חדשה מכונית
-  nuevo automovil

What if both languages are MRLs?

Example:

- Spanish and Hebrew have **adjective-noun** agreement

– new shirt

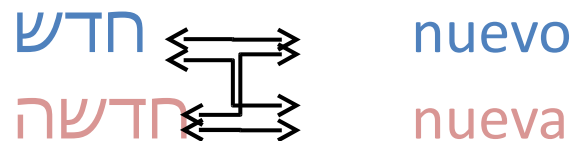
- חולצה חדשה
-  nueva camisa

– new computer

- מחשב חדש
-  nueva computadora

– new car

- מכונית חדשה
-  nuevo automovil



What if both languages are MRLs?

Ex:

- Many-to-many mapping
- Correct form still depends on external information
- More chances for error
- Acquiring all the pairs from parallel corpora is

– ^{harder} new shirt

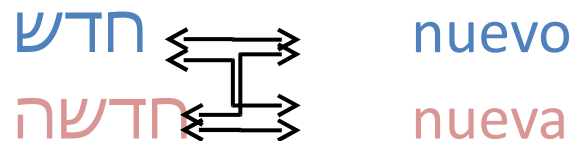
- חדשה חולצה
-  nueva camisa

– new computer

- חדש מחשב
-  nueva computadora

– new car

- חדשה מכונית
-  nuevo automovil



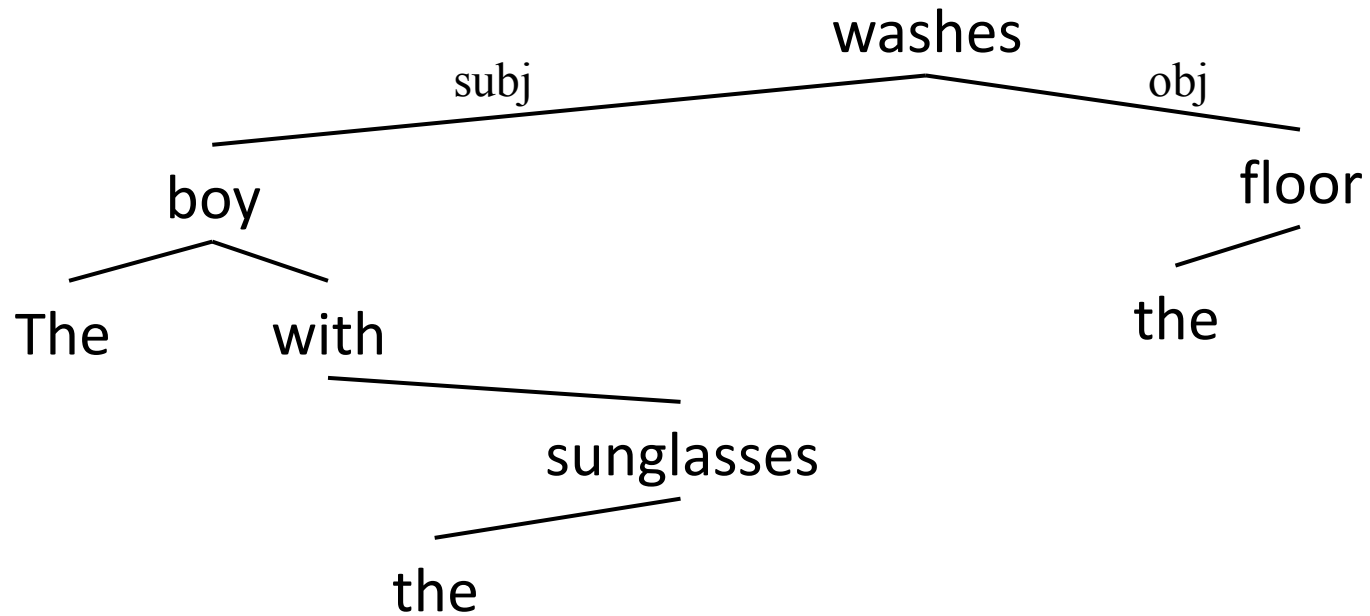
Phrase-tables and n-grams
still can't do it

Must consider (at least) syntax

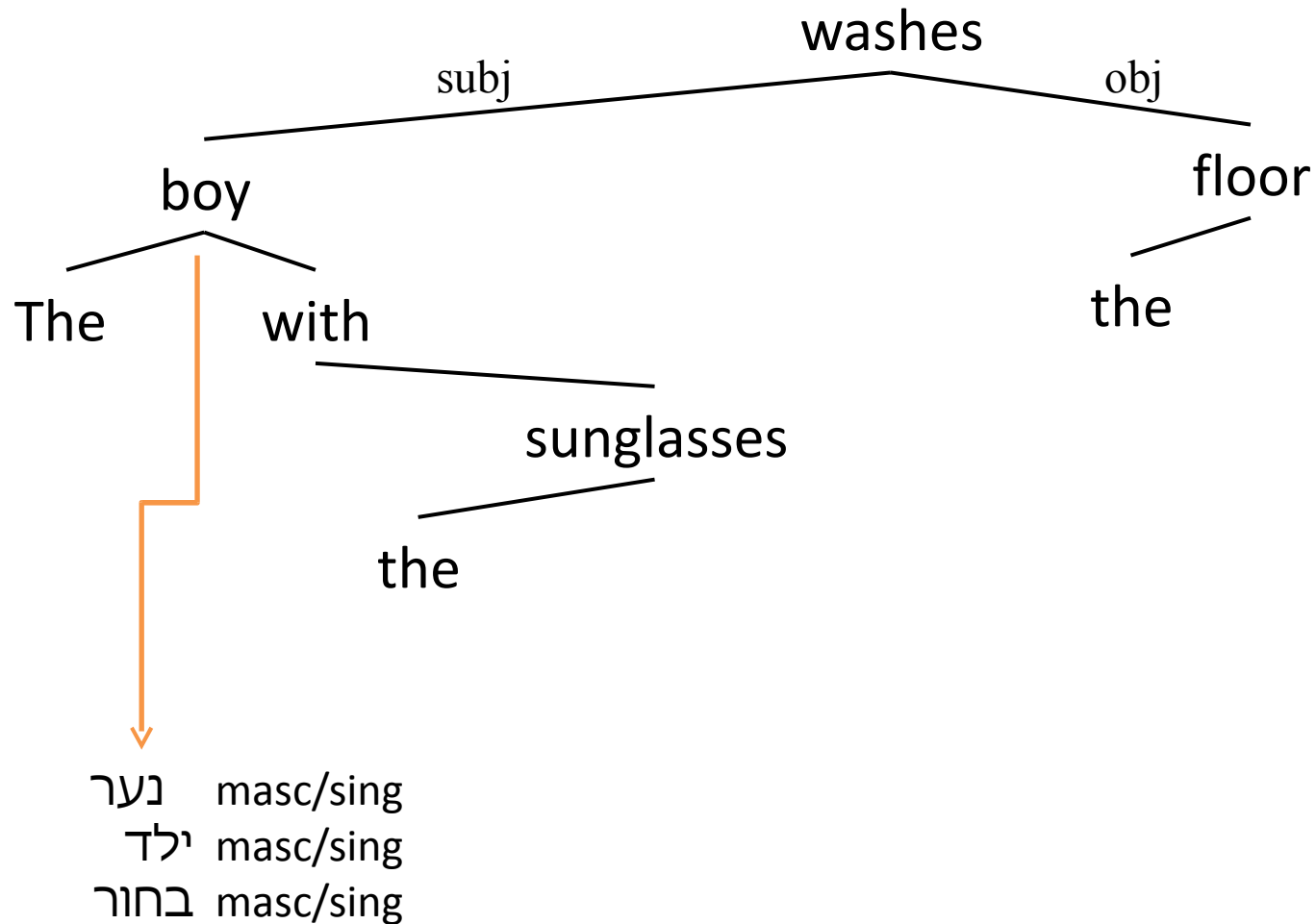
When Translating into an MRL:

- MT systems must be aware of gender/number
- Should have a notion of agreement
- Use syntax to enforce agreement

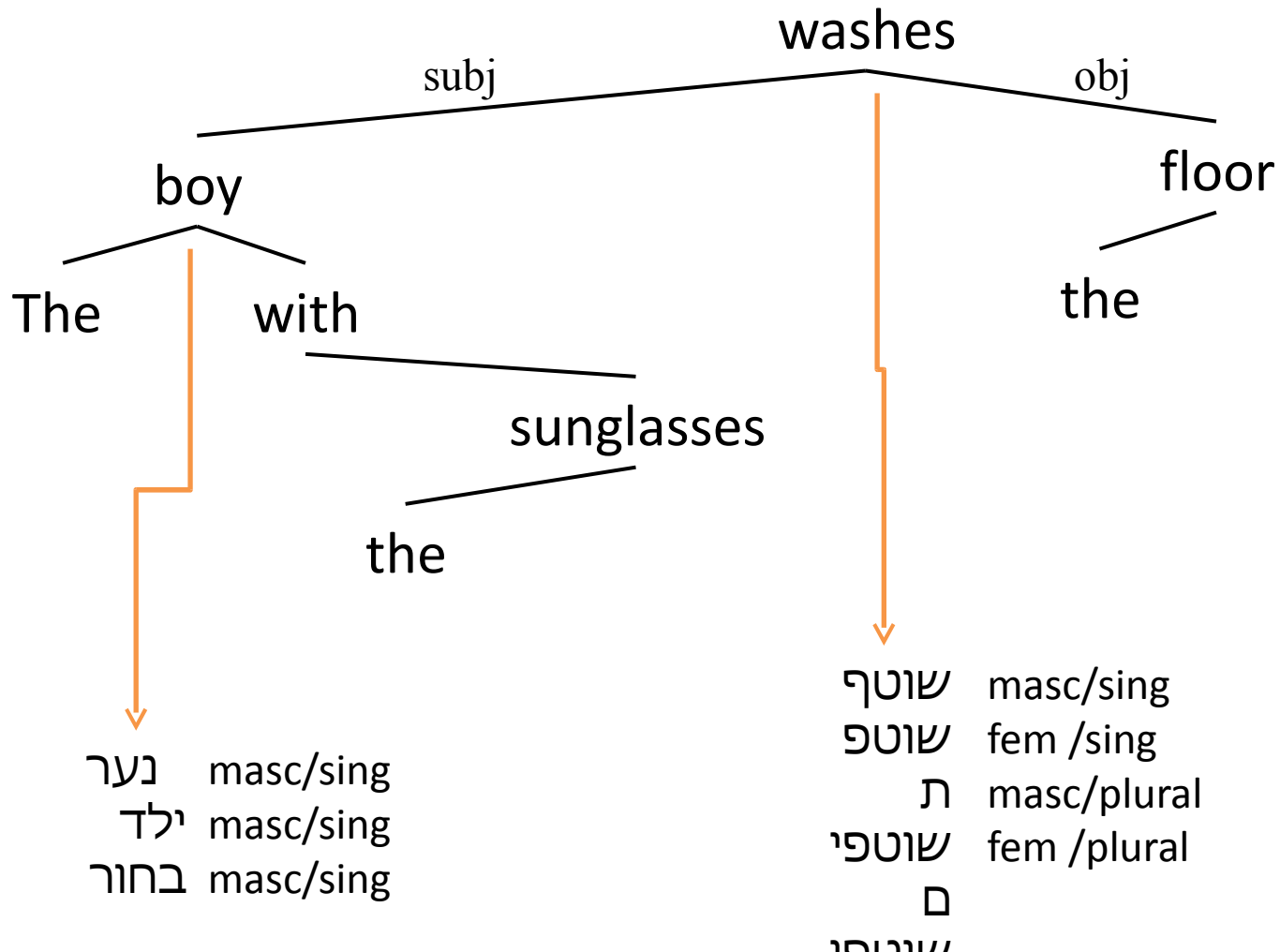
Source-side Syntax



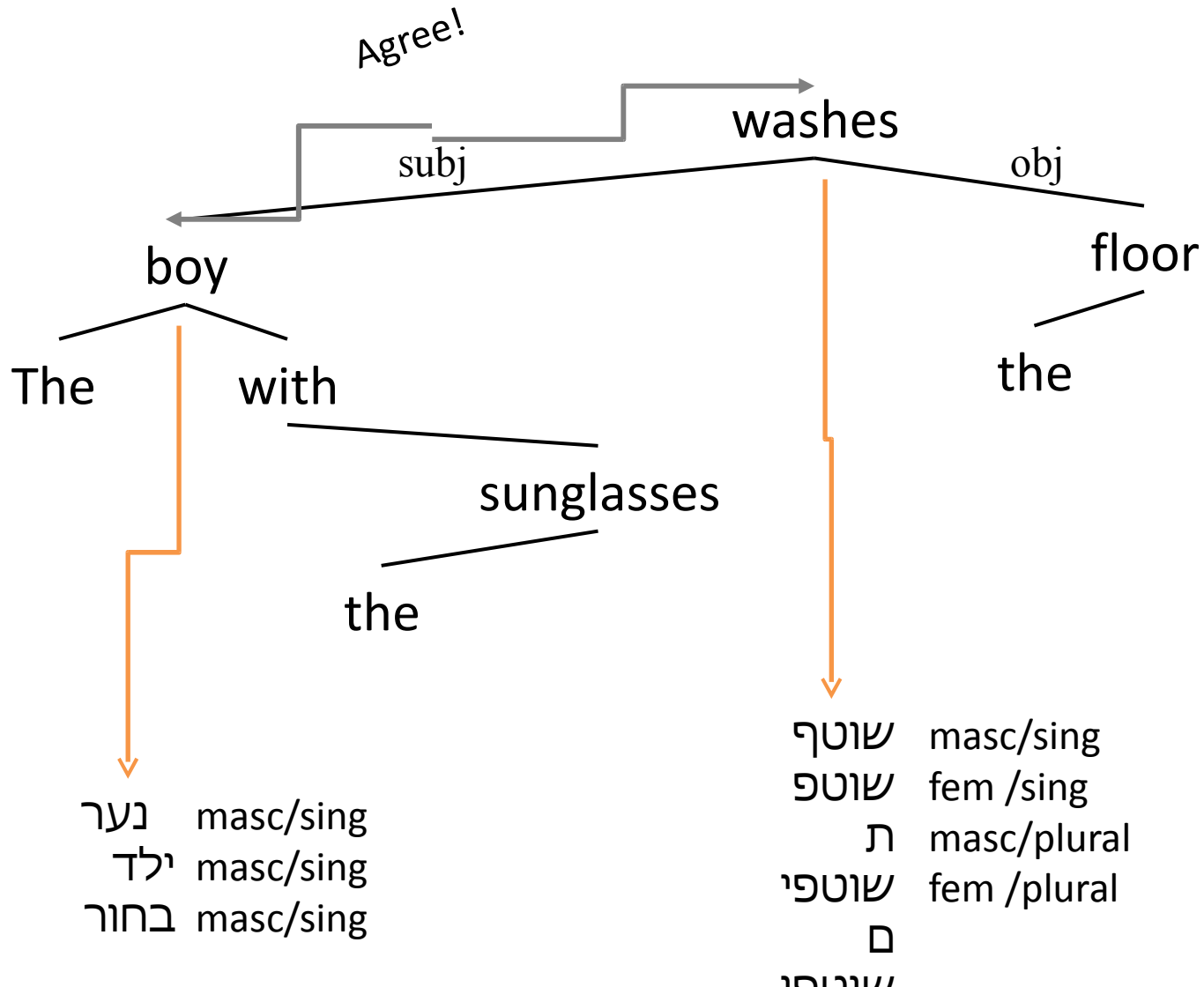
Source-side Syntax



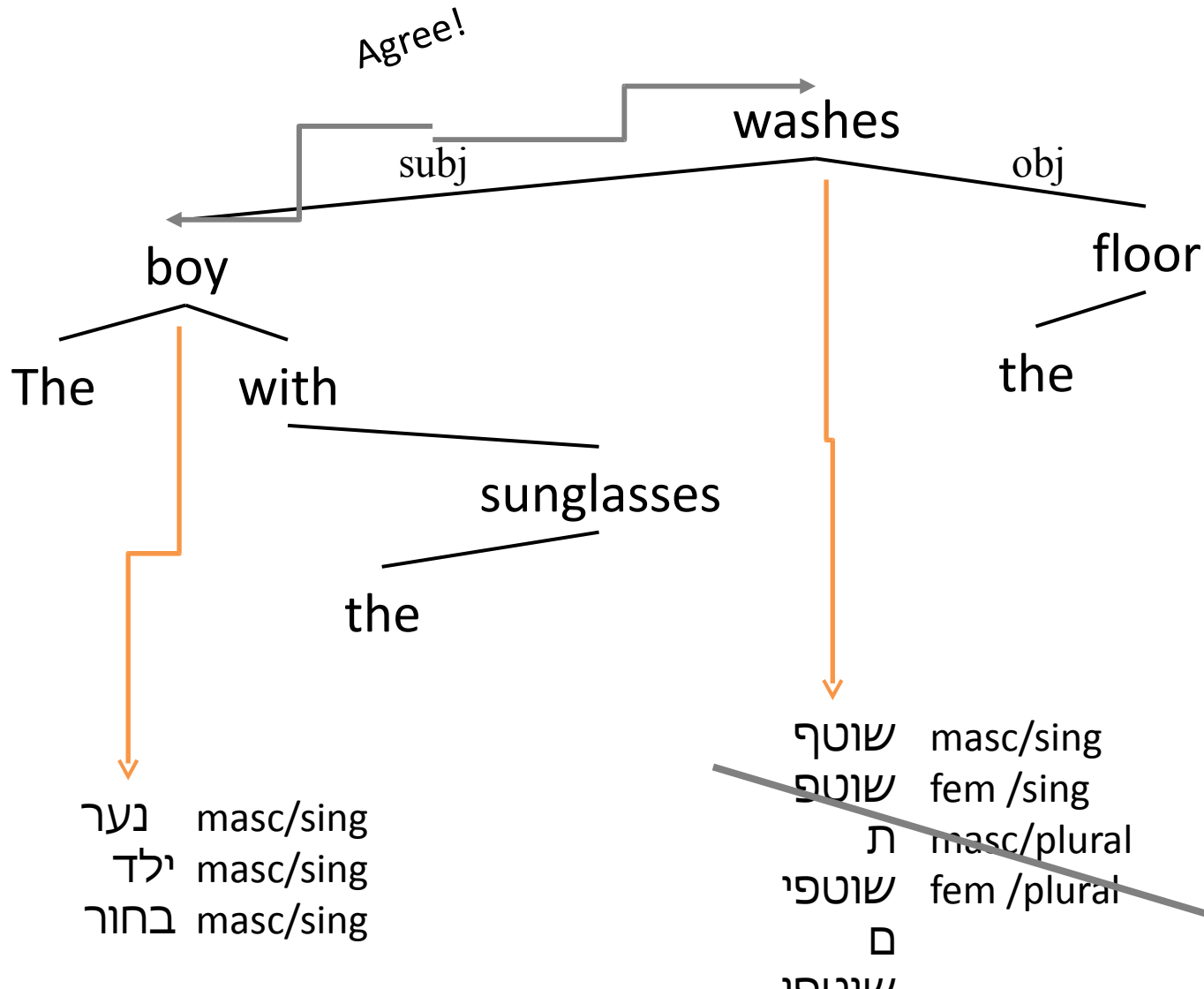
Source-side Syntax



Source-side Syntax



Source-side Syntax



Source-side Syntax

washes

Problems:

- How to obtain gender/number information?
- How to decode efficiently?
- Agreement behavior is not always that simple

the

נער masc/sing
ילד masc/sing
בחור masc/sing

שׁוֹטֵף masc/sing
~~שׁוֹטֵפָה fem /sing~~
~~שׁוֹטְפִים masc/plural~~
שׁוֹטְפֵי fem /plural
ם

Target-side Syntax

xLNT transducers can model agreement

Target-side Syntax

xLNT transducers can model agreement

NN_{masc/sing} (יָלֵד) boy

VB_{masc/sing} (שֹׁטֵף) washes

VB_{fem/sing} (שֹׁטֶפֶת) washes

NP_{masc/sing}(x0:DT x1:NN_{masc/sing} x2:PP) x0 x1

x2

VP_{masc/sing} (x0:VB_{masc/sing} x1:NP) x0

x1

S_{masc/sing}(x0:NP_{masc/sing} x1:VP_{masc/sing}) x0 x1

x0 x1

Target-side Syntax

xLNT transducers can model agreement

NNmasc/sing (יֵלֵד) □ boy
VBmasc/sing (שׁוֹטֵף) □ washes
VBfem/sing (שׁוֹטֶפֶת) □ washes

**Gender and number
information
encoded in the lexical rules**

NPmasc/sing(x0:DT x1:NNmasc/sing x2:PP)

□ x0 x1 x2

VPmasc/sing (x0:VB masc/sing x1:NP)

□ x0 x1

Smasc/sing(x0:NPmasc/sing x1:VPmasc/sing)

□ x0 x1

Target-side Syntax

xLNT transducers can model agreement

NNmasc/sing (יֵלֵד) □ boy
VBmasc/sing (שׁוֹטֵף) □ washes
VBfem/sing (שׁוֹטֶפֶת) □ washes

**Gender and number
information
encoded in the lexical rules**

NPmasc/sing(x0:DT x1:NNmasc/sing x2:PP)

□ x0 x1 x2

VPmasc/sing (x0:VB masc/sing x1:NP)

□ x0 x1

Smasc/sing(x0:NPmasc/sing x1:VPmasc/sing)

□ x0 x1

**Agreement information
encoded in the grammar**

Target-side Syntax

xLNT transducers can model agreement

Problems:

- How to obtain gender/number information?
- Grammar is going to be huge (can we make it smaller?)
- How are we going to obtain the grammar?

efficiently encoding morphological processes in a treebank grammar ? an open research question

On The Parsing Side of Things

- Most work on parsing MRLs:
 - consider morphology to be a lexicon-level issue
 - Many inflections \Rightarrow high OOV rate

→ **Modest benefits to parsing accuracy**

→ **PCEGLA still better \Rightarrow**

On The Parsing Side of Things

- Most work on parsing MRLs:
 - consider morphology to be a lexicon-level issue
 - Many inflections \Rightarrow high OOV rate
 - Ignoring morphology at syntax-level

PCFG A works frustratingly well

→ **Modest benefits to parsing accuracy**

→ **PCFG A still better \Rightarrow**

On The Parsing Side of Things

- Most work on parsing MRLs:
 - consider morphology to be a lexicon-level issue
 - Many inflections → high OOV rate
 - Ignoring morphology at syntax-level
 - **PCFGLA works frustratingly well**
- Recently:
 - **Modest benefits to parsing accuracy**
 - **PCFGLA still better** →

On The Parsing Side of Things

- Most work on parsing MRLs:
 - consider morphology to be a lexicon-level issue
 - Many inflections → high OOV rate
 - Ignoring morphology at syntax-level
 - **PCFGLA works frustratingly well**
- Recently:
 - Smarter modeling of morphology at syntax level
 - **Modest benefits to parsing accuracy**
 - **PCFGLA still better** →

On The Parsing Side of Things

- Most work on parsing MRLs:
 - consider morphology to be a lexicon-level issue
 - Many inflections \Rightarrow high OOV rate
 - Ignoring morphology at syntax-level
 - **PCFGLA works frustratingly well**
- Recently:
 - Smarter modeling of morphology at syntax level
 - Using morphological agreement to improve
 - \rightarrow **Modest benefits to parsing accuracy**
 - \rightarrow **PCFGLA still better \Rightarrow**

On The Parsing Side of Things

- Most work on parsing MRLs:
 - consider morphology to be a lexicon-level issue
 - Many inflections \Rightarrow high OOV rate
 - Ignoring morphology at syntax-level
 - **PCFGLA works frustratingly well**
- Recently:
 - Smarter modeling of morphology at syntax level
 - Using morphological agreement to improve parsing
 - \rightarrow **Modest benefits to parsing accuracy**
 - \rightarrow **PCFGLA still better \Rightarrow**

On The Parsing Side of Things

- Most work on parsing MRLs
 - consider morphology to be a nuisance
 - Many inflections \Rightarrow high OOV rate
 - Ignoring morphology at syntax-level
 - **PCFGLA works frustratingly well**
- Recently:
 - Smarter modeling of morphology at syntax level
 - Using morphological agreement to improve parsing
 - \rightarrow **Modest benefits to parsing accuracy**
 - \rightarrow **PCFGLA still better \Rightarrow**

PCFGLA
is not
modeling agreement

Rebels without a cause?

- Syntax-based MT:
 - Neat!
 - Only marginally better than phrase-based

Rebels without a cause?

- Syntax-based MT:

- Neat!

- Only marginally better than phrase-based

English grammaticality is relatively easy to capture using local information

Rebels without a cause?

- Syntax-based MT:

- Neat!

- Only marginally better than phrase-based

- Syntax-level morphology in parsing:

- Neat!

- Only marginally better than ignoring it

English grammaticality is relatively easy to capture using local information

Rebels without a cause?

English grammaticality is

- Syntax-based
 - Neat!
 - Only marginally better than ignoring it
- I should work harder.
- Not many agreement mistakes to begin with.
- Agreement is a generation issue more than a parsing one.
- Syntax-less
 - Neat!
 - Only marginally better than ignoring it

Rebels without a cause?

- Syntax-based MT:
 - Neat!
 - Only marginally better than phrase-based
- Syntax-level morphology in parsing:
 - Neat!
 - Only marginally better than ignoring it

Both are crucial for translating *into* MRLs!

To Conclude

- Translating into MRLs brings new challenges
- Syntax is crucial
 - If **you are not** looking into syntax, you should!
 - If **you are** looking into syntax – look deeper!
- Plenty of interesting work to be done

To Conclude

- Translating into MRLs brings new challenges
- Syntax is crucial
 - If **you are not** looking into syntax, you should!
 - If **you are** looking into syntax – look deeper!
- Plenty of interesting work to be done
 - **Finishing up my phd on parsing**
Looking for a postdoc position for next year