DIRETRA, a customizable direct translation system: first sketches

Alëna Aksënova **&** **Marina Ermolaeva**

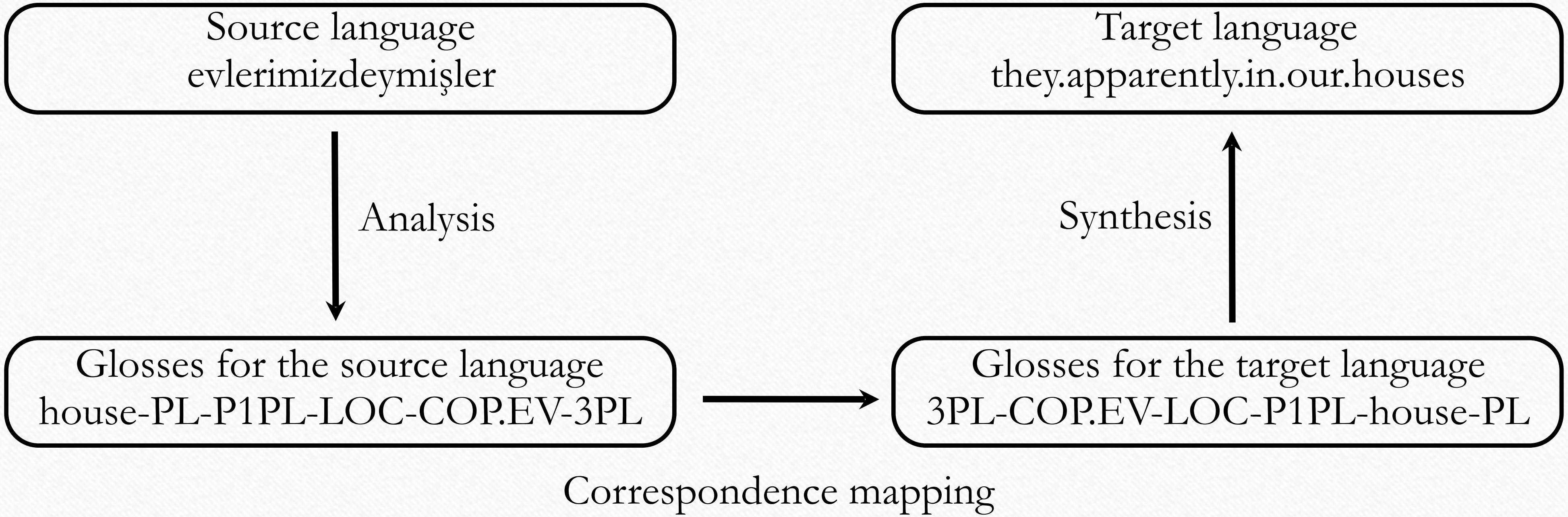
loisetoil@gmail.com marinkaermolaeva@mail.ru

Translata II, 30 October, 2014

The Zen of Diretra

- Diretra = direct translation system with morphology.
- Diretra ≠ translation system.
- Linguistic generalizations are our friends.
- If it exists, it should be taken into account.
- First general, then specific.
- Errors are intolerable.
- K.I.S.S.

Bird’s-eye View



Source language
evlerimizdeymişler

Analysis

Glosses for the source language
house-PL-P1PL-LOC-COP.EV-3PL

Correspondence mapping

Glosses for the target language
3PL-COP.EV-LOC-P1PL-house-PL

Synthesis

Target language
they.apparently.in.our.houses

Turkish Challenges: Multiple Allomorphs

- Vowel harmony:
  - palatal (±front)
  - labial (±rounded)
  - up to 4 allomorphs
- Assimilation and avoiding hiatus:
  - voiced/voiceless
  - vowel/consonant

	unrounded	rounded
back	(1) baş-**ın** head-P2SG	(2) kol-**un** arm-P2SG
front	(3) ev-**in** house-P2SG	(4) göz-**ün** eye-P2SG

Turkish Challenges: Complexities Within Roots

- Alternations in roots:
  - voiced/voiceless final consonant
  - vowel/∅
  - single/double consonant
- Compounds written in one word
- Exceptions to harmony:
  - within stems and compounds
  - at the stem-suffix boundary

(5) şehir
city

(6) şehr-i
city-ACC

(7) kız-arkadaş
girl-friend

(8) sarı-humma
yellow-fever

(9) hal-in
condition-P2SG

Turkish Challenges: Morphology

- Main morphological categories of nouns:
  - number;
  - possession;
  - case.
- Nominals can subsequently form predicates and adverbials.

(10)

STEM	NUM	POSS	CASE	COPULA	PERS+NUM
ev	-ler	-imiz	-de	-ymiş	-ler
house	-PL	-P1PL	-LOC	-COP.EV	-3PL

Apparently, they are/were in our houses.

(11)

STEM	NUM	POSS	CASE	ADV
sokak	-∅	-∅	-ta	-yken
street	-SG	-NPS	-LOC	-while

while in the street

Turkish Challenges: Morphology

- The suffix -ki attaches to nominals in locative or genitive case;
- Word forms with -ki can receive nominal suffixes themselves;
- Recursion: -ki can attach to locative and genitive forms that already contain a -ki.
- LOC-ki and GEN-ki have different properties (Hankamer 2004);

(12) raf-ta-ki
shelf-LOC-KI1
the one on the shelf

(13) Hasan-ın-ki
Hasan-GEN-KI2
Hasan's

(14) ev-de-ki-ler-in-ki
house-LOC-KI1-PL-GEN-KI2
the one belonging to those at home

Parser: Goals

- Adaptability;
- Good results with limited resources:
  - analyze morphology even if the stem is unknown
  - right-to-left processing
- Dealing with complex cases:
  - stems with special properties
  - recursive affixes
  - compounds

Parser: The Main Idea

- A hybrid approach;
- The main unit is a **slot** – a part of the affix chain with a fixed sequence order (or orders);
- For each slot we list all category sequences that can fill it.

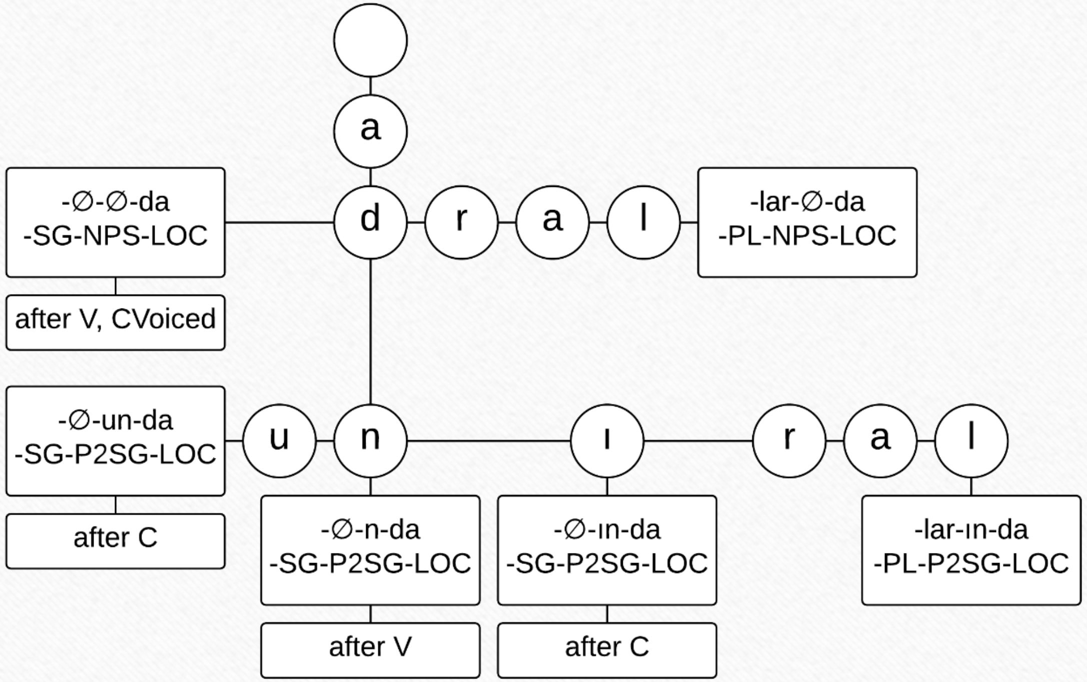
Category	**Sequences**	**Slot**
Number	-NUM-POSS-CASE	Nominal inflection
Possession		
Case		
Question particle	-(Q)-COP.PRS-PERS -(Q)-COP.PST-PERS -ADV …	Nominal verb suffixes
Copula		
Person & number markers		
Adverbial markers		
…	…	…

Parser: The Main Idea

- Currently, the set of slots is fixed:
  - modifier stem
  - main stem
  - noun inflection
  - loop within noun inflection
  - nominal verb suffixes
- The number and order(s) of categories within slots can be changed easily.

Parser: Data Representation

- For each slot, all possible affix chains are constructed;
- The lexicon and affix sequences are inverted and stored in a set of tries;
- Phonological variants of a same stem or affix are treated as separate entries.



-∅-∅-da
-SG-NPS-LOC

after V, CVoiced

-lar-∅-da
-PL-NPS-LOC

-∅-un-da
-SG-P2SG-LOC

after C

-∅-n-da
-SG-P2SG-LOC

after V

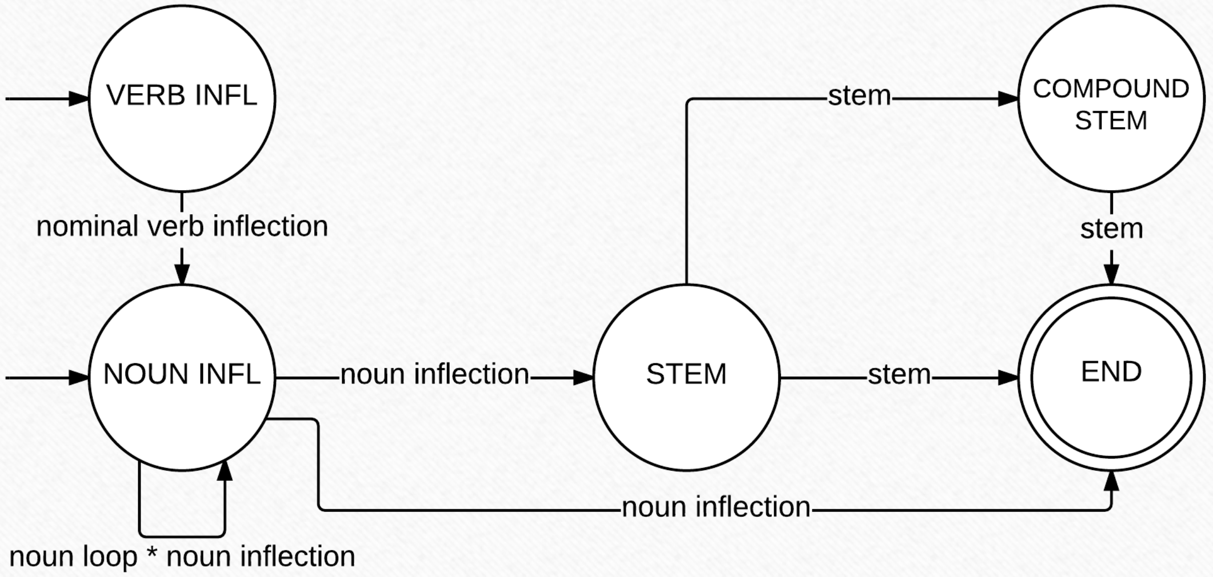
-∅-ın-da
-SG-P2SG-LOC

after C

-lar-ın-da
-PL-P2SG-LOC

Parser: the Algorithm

- The transitions between slots are performed via a FSM;
- Right-to-left processing;
- Possible decisions:
  - single known stem;
  - compound (two known stems);
  - unknown stem.



Correspondence Mapping: The Main Idea

- What is syntax in some languages is morphology in the others;
- Turkic languages have rich morphology;
- English morphology is much less complicated;
- Current task: complex-to-simple.

- The Mirror Principle (Baker 1985):

Morphological derivations must directly reflect syntactic derivations (and vice versa).

Correspondence Mapping: The Main Idea

- Affix sequences for the target language “mirror” those for the source language;
- Affixes that are represented as morphemes and attach to the stem in the target language stay in place.

(15) ada-∅-m-∅
island-SG-P1SG-NOM
3PL-COP.PRS-NOM
my.island

(16) çocuk-lar-∅-ın-ki-ler-∅-∅
child-PL-NPS-GEN-KI2-PL-NPS-NOM
NOM-NPS-PL-KI2-GEN-NPS-child-PL
ones.owned.by.children

Correspondence Mapping: Minor Details

- Auxiliary movement in general questions (target language):
- The presence of a question marker triggers the movement of copulas to the leftmost position.

(17) çocuk-lar-∅-∅-mı-∅-yız
child-PL-NPS-NOM-Q-**COP.PRS-1PL**
COP.PRS-1PL-NOM-NPS-child-PL-Q
are.we.children.?

Synthesis: Replacement Rules

- 5 types of rules:
  - simple replacement (“DAT” → “to”, “P1SG” → “my”)
  - morphology-driven replacement (“COP.PRS” → “am” if 1SG, “is” if 3SG, “are” elsewhere)
  - phonology-driven replacement (“PL” → “ies” if __Cy, → “PL” → “ves” if __f etc.)
  - application of irregular forms (“deer” + “PL” → “deer”)
  - statistics-based replacement (“LOC” → “in” or “on” or “at”)

Synthesis: Simple Replacement

- “DAT” → “to”
- “1SG” → “I”, “1PL” → “we”, ...
- “P1SG” → “my”, “P1PL” → “our”, ...

(18) adam-∅-∅-a
man-SG-NPS-DAT
DAT-NPS-man-SG
to.man

(19) elma-∅-m-∅
apple-SG-P1SG-NOM
NOM-P1SG-apple-SG
my.apple

Synthesis: Morphology-driven Replacement

- Implementation of agreement;
- “ACC” → “the”, but no “the” in possessives and proper names
  - DOM in Turkic languages, Lyutikova&Pereltsvaig (2013)

(20) çocuk-∅-∅-∅-mu-∅-yum
child-SG-NPS-NOM-Q-COP.PRS-1SG
COP.PRS-1SG-NOM-NPS-child-SG-Q
am.I.child.?

(21) arkadaş-∅-∅-ı
friend-SG-NPS-ACC
ACC-NPS-friend-SG
the.friend

(22) arkadaş-∅-ım-ı
friend-SG-P1SG-ACC
ACC-P1SG-friend-SG
my.friend

Synthesis: Word Form Generation

- Phonological rules are applied to build regular plural forms;
- A list of irregular plurals is used in order to generate irregular plural forms correctly.

(23) karı-lar-∅-∅
wife-PL-NPS-NOM
NOM-NPS-wife-PL
wives

(24) sihirbaz-lar-∅-∅
witch-PL-NPS-NOM
NOM-NPS-witch-PL
witches

(25) geyik-ler-∅-∅
deer-PL-NPS-NOM
NOM-NPS-deer-PL
deer

(26) eksen-ler-∅-∅
axis-PL-NPS-NOM
NOM-NPS-axis-PL
axes

Synthesis: Statistics-based Replacement

- The target language lacks a morphological locative case;
- For each target language noun there is a locative preposition that is used with it most often;
- We calculate frequencies to determine the best replacement for the "LOC" gloss.

(27) ev-∅-∅-de
house-SG-NPS-LOC
LOC-NPS-house-SG
in.house

(28) okul-∅-∅-da
school-SG-NPS-LOC
LOC-NPS-school-SG
at.school

Future Work

- Deeper:
  - finite and nonfinite verb forms
  - other parts of speech
  - derivational morphology
- Wider:
  - other Turkic languages
  - other Altaic languages

References

- Baker M. (1985). The Mirror Principle and Morphosyntactic Explanation // Linguistic Inquiry 16. 373-416.
- Çöltekin Ç. (2010). A Freely Available Morphological Analyzer for Turkish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias, eds., 'LREC', European Language Resources Association.
- Eryiğit G., Adalı E. (2004). An Affix Stripping Morphological Analyzer for Turkish // IASTED International Multi-Conference on Artificial Intelligence and Applications. Innsbruck, Austria, pages 299−304.
- Göksel A., Kerslake C. (2005). Turkish: A Comprehensive Grammar.
- Hankamer J. (1986). Finite state morphology and left-to-right phonology // Proceedings of the Fifth West Coast Conference on Formal Linguistics, Stanford, CA, pages 29−34.

References

- Hankamer J. (2004). Why there are two ki's in Turkish // Imer and Dogan, eds., Current Research in Turkish Linguistics, Eastern Mediterranean University Press, 13-25.
- Kornfilt J. (1996). On copular clitic forms in Turkish. ZAS Papers in Linguistics 6, 96-114.
- Kornfilt J. (1997). Turkish. London and New York: Routledge.
- Lewis G. (1967). Turkish Grammar. Oxford: Oxford University Press.
- Lyutikova E., Pereltsvaig A. (2013). Elucidating nominal structure in articleless languages: A case study of Tatar // Proceedings of 39th Berkeley Linguistic Society Meeting. — Berkeley.