Deep-syntax TectoMT for English-Spanish MT

Gorka Labaka, Oneka Jauregi, Arantza Díaz de Ilarraza, Michael Ustaszewski, Nora Aranberri and Eneko Agirre

IXA Group
University of the Basque Country, Spain

Deep Machine Translation Workshop, Prague, September 3-4, 2015

Outline

- TectoMT architecture
- Development of a new language pair (English - Spanish)
  - Analysis
  - Transfer
  - Synthesis
- Evaluation
- Conclusions

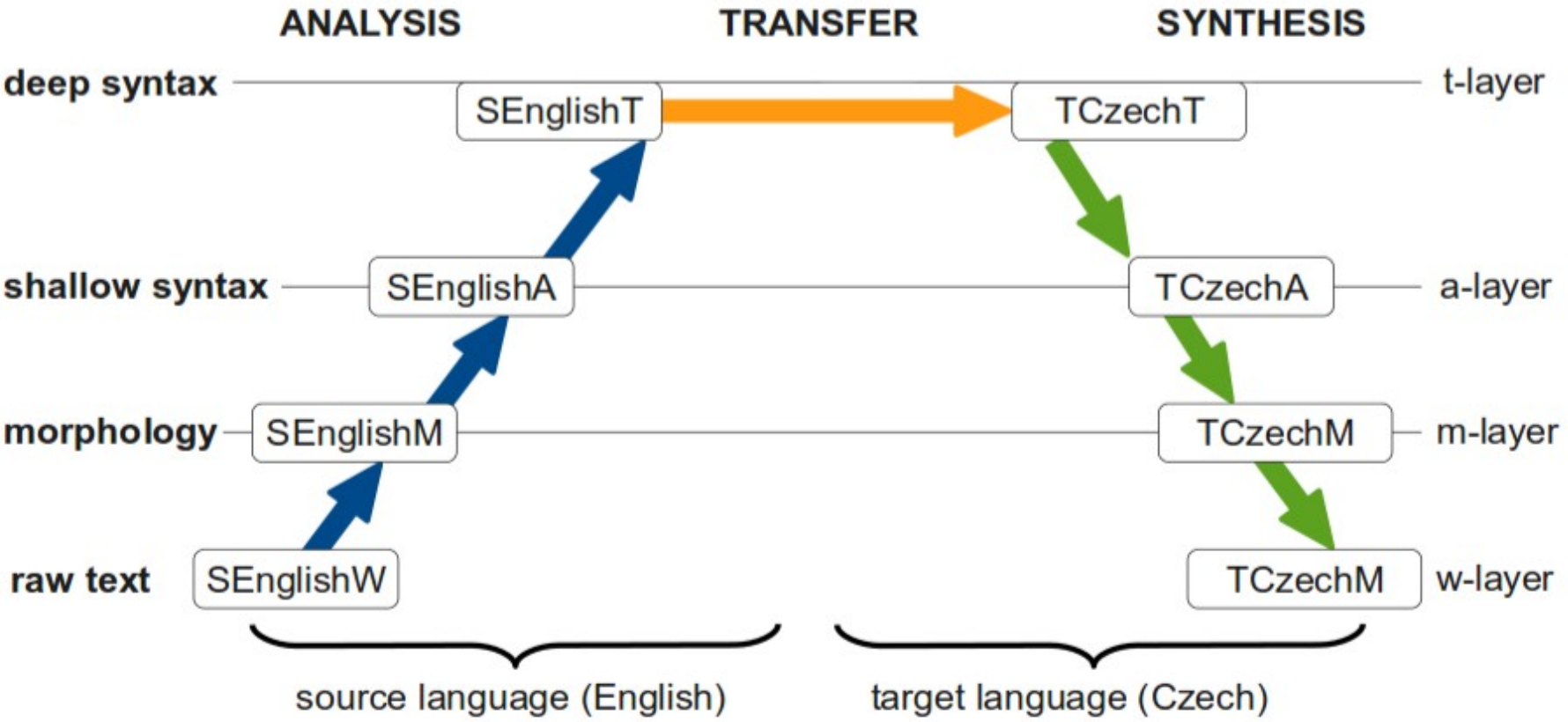
Tecto layers

- TectoMT
  - transfer-based system which works at the deep tectogrammatical level
  - combines linguistic knowledge and statistical techniques, particularly during transfer
  - originally developed for the English-Czech language direction
- Stratification approach
  - Morphological layer
  - Analytical layer (shallow-syntax dependency tree)
  - Tectogrammatical layer (deep-syntax dependency tree)

Tectogrammatical layer

- Only autosemantic nodes are keeped
- Functional words represented by attributes
- Each t-node consists on:
  - Tectogrammatical lemma
  - Functor: semantic values of syntactic dependency relations (causal adjunt, actor, effect, etc.)
  - Grammatemes: semantically oriented morphological categories (tense, number, modality, etc.)
  - Formemes: values of the morphosyntactic form in the surface sentence (subject, direct object, etc.)

TectoMT architecture



Tecto blocks and scenarios

- Blocks: reusable components of NLP subtasks that can be listed in a specific sequence, that is, rules to define, set, change and move node-information in/across the layers
- Scenarios: specific sequences of blocks to be applied to relevant data
- TectoMT includes over a thousand blocks:
  - 224 blocks specific for English
  - 237 for Czech
  - 57 for English-to-Czech transfer
  - 129 for other languages
  - 467 language-independent

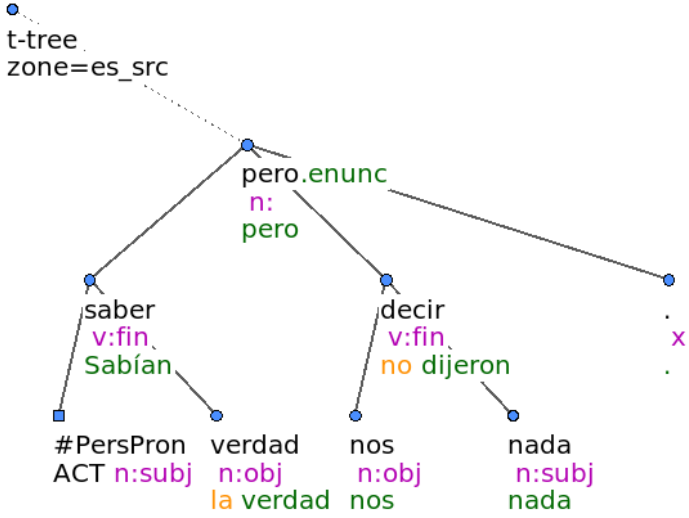
Developing a new pair

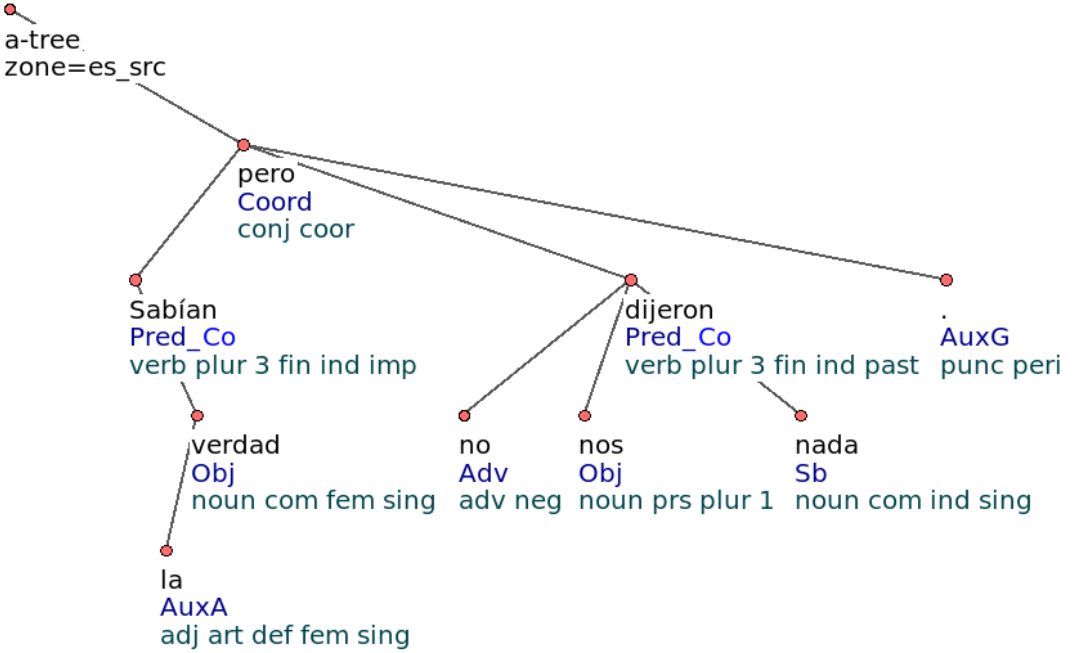
- We set to port the TectoMT system to work for the English-Spanish language pair in both directions.
  - English analysis and synthesis ready to use
  - Our focus: Spanish analysis and synthesis, and transfer stages
- TectoMT is integrated within Treex
  - Modules divided into language-specific and language independent blocks

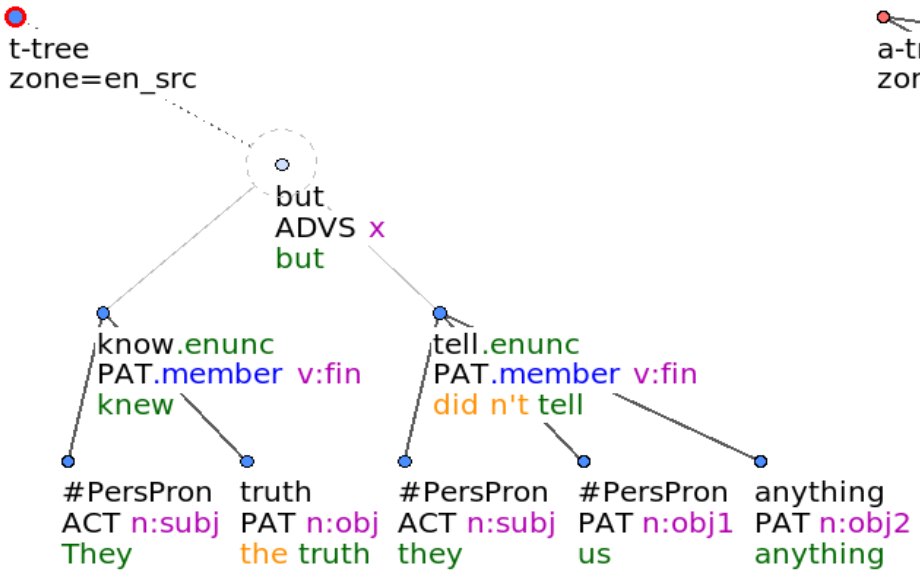
Analysis

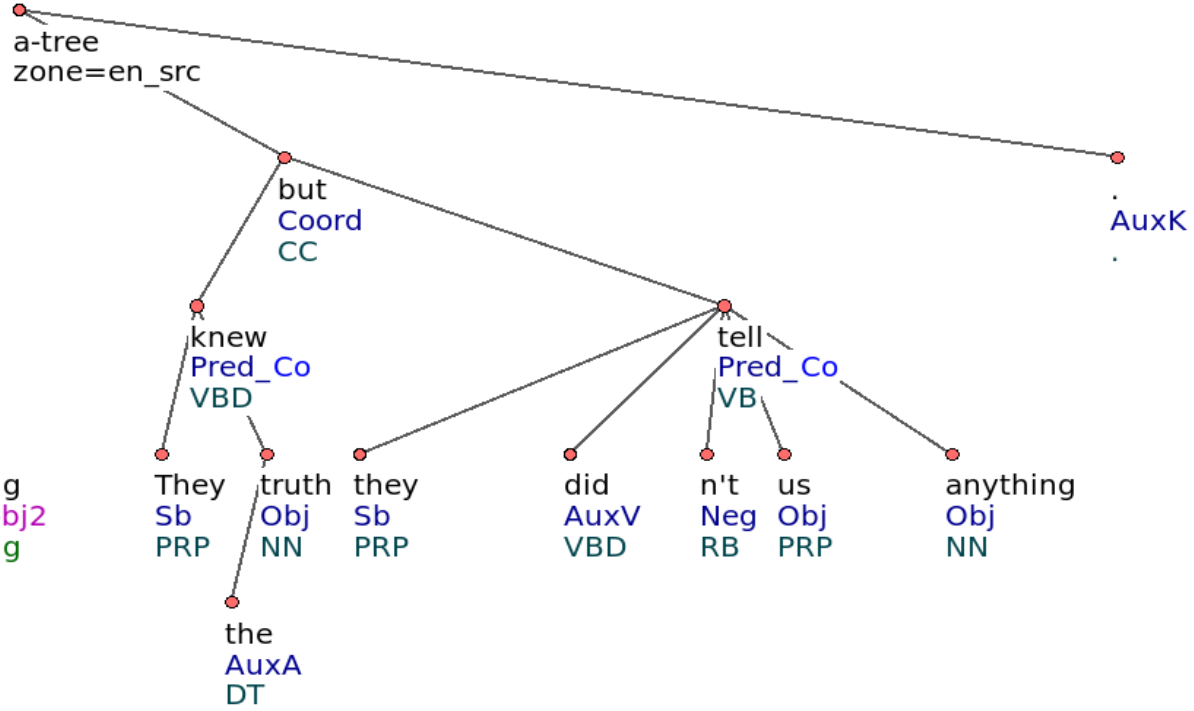
- From raw text to tecto-level
- English analysis solved
- Spanish analysis
  - Tokenization and sentence splitting: adapted modules in Treex
  - Lemmatisation and POS: integration of ixa-pipes tools (pos) in Treex
  - Dependency parsing: integration of ixa-pipes tools (srl) in Treex
    - Tagset compatibility: from AnCora to Interset
  - Spanish blocks:

Block type	Number
Language-independent blocks	11
Adapted blocks	4
New language-specific blocks	1


t-tree
zone=es_src
pero.enunc
n:
pero
saber
v:fin
Sabían
decir
v:fin
no dijeron
.
x
.
#PersPron
ACT n:subj
verdad
n:obj
la verdad
nos
n:obj
nos
nada
n:subj
nada


a-tree
zone=es_src
pero
Coord
conj coor
Sabían
Pred_Co
verb plur 3 fin ind imp
dijeron
Pred_Co
verb plur 3 fin ind past
.
AuxG
punc peri
verdad
Obj
noun com fem sing
no
Adv
adv neg
nos
Obj
noun prs plur 1
nada
Sb
noun com ind sing
la
AuxA
adj art def fem sing


t-tree
zone=en_src
but
ADVS x
but
know.enunc
PAT.member v:fin
knew
tell.enunc
PAT.member v:fin
did n't tell
#PersPron
ACT n:subj
They
truth
PAT n:obj
the truth
#PersPron
ACT n:subj
they
#PersPron
PAT n:obj1
us
anything
PAT n:obj2
anything


a-tree
zone=en_src
but
Coord
CC
.
AuxK
.
knew
Pred_Co
VBD
tell
Pred_Co
VB
They
Sb
PRP
truth
Obj
NN
the
AuxA
DT
they
Sb
PRP
did
AuxV
VBD
n't
Neg
RB
us
Obj
PRP
anything
Obj
NN

Transfer

- Statistical transfer dictionary
  - trained on parallel corpora analyzed up to the t-level in both languages
    - lemmas, formemes and grammatemes
  - for each t-lemma and formeme in a source t-tree, the translation model assigns a score to all possible translations observed in the training data
  - probability estimate calculated as a linear combination of
    - Discriminative TM
    - Dictionary TM
- Static manual dictionary (priority resource)
  - Microsoft Terminology Collection - 22,475 entries

Transfer

- Blocks for grammateme equivalences
  - linguistically abstract, usually paralleled in the target language
- rules are inherently language-specific
- 5 blocks for English-to-Spanish direction:
  - lack of gender in English nouns (necessary in Spanish);
  - differences in definiteness and articles;
  - differences in structures such as “There is...” and relative clauses.

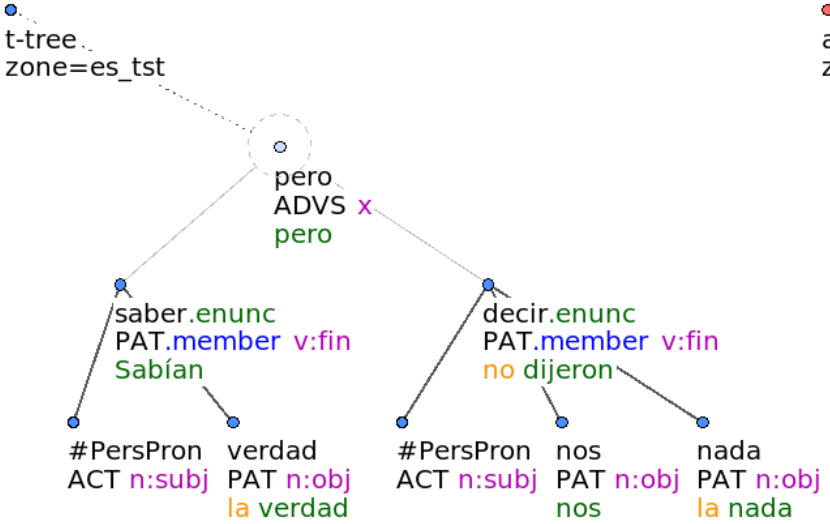
Synthesis

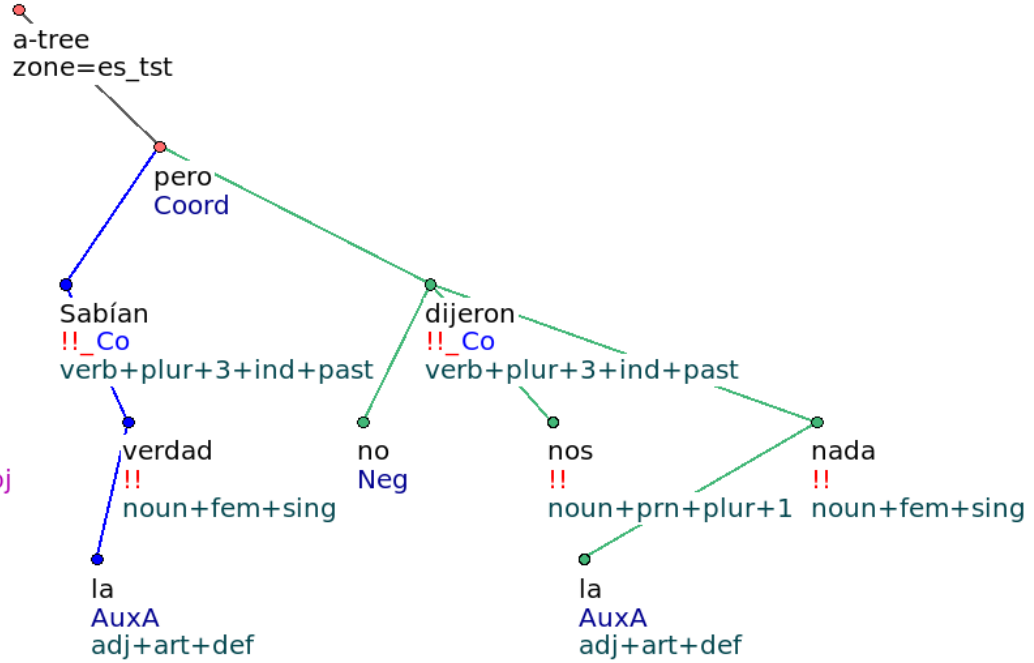
- From tecto-level to raw text
- English synthesis solved
- Spanish synthesis
  - Transform the t-tree into an a-tree
  - Transform the a-tree into word forms
  - Polish the output

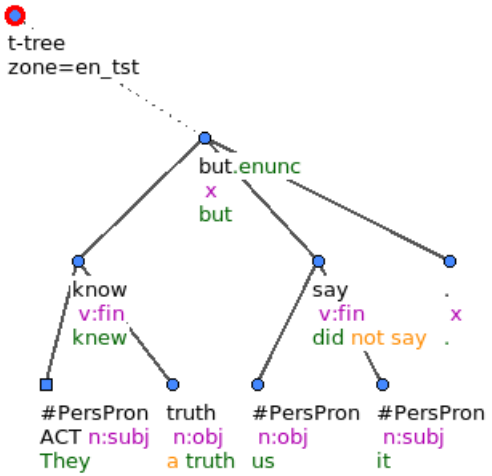
Block type	Number
Language-independent blocks	9
Adapted blocks	12
New language-specific blocks	3

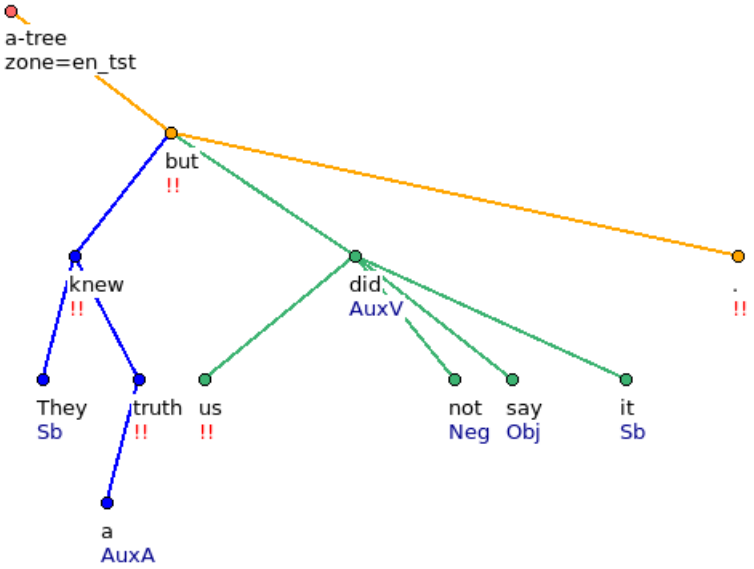
Synthesis

- Transform the t-tree into an a-tree:
  - fill in morphological attributes that will be needed in the second step
  - add function words where necessary
  - remove superfluous nodes
  - add punctuation nodes
- Transform the a-tree into word forms
  - new Spanish models in Flect (statistical morphological generator)
  - corpus: subset of morphologically annotated (530K tokens)
- Polish the output: detokenization, contractions, ...


t-tree
zone=es_tst
pero
ADVS x
pero
saber.enunc
PAT.member v:fin
Sabían
decir.enunc
PAT.member v:fin
no dijeron
#PersPron
ACT n:subj
verdad
PAT n:obj
la verdad
#PersPron
ACT n:subj
nos
PAT n:obj
nos
nada
PAT n:obj
la nada


a-tree
zone=es_tst
pero
Coord
Sabían
!!_Co
verb+plur+3+ind+past
dijeron
!!_Co
verb+plur+3+ind+past
verdad
!!
noun+fem+sing
no
Neg
nos
!!
noun+prn+plur+1
nada
!!
noun+fem+sing
la
AuxA
adj+art+def
la
AuxA
adj+art+def


t-tree
zone=en_tst
but.enunc
x
but
know
v:fin
knew
say
v:fin
did not say
.
x
.
#PersPron
ACT n:subj
They
truth
n:obj
a truth
#PersPron
n:obj
us
#PersPron
n:subj
it


a-tree
zone=en_tst
but
!!
knew
!!
did
AuxV
.
!!
They
Sb
truth
!!
us
!!
not
Neg
say
Obj
it
Sb
a
AuxA

Evaluation

- Compared systems:
  - PBSMT (Moses)
    - Features: mGiza, SRILM
    - Corpora:
      - Bilingual: europarl (~2M sentences)
      - Monolingual: europarl (~2M sentences)
    - Tuning: 1,000 IT-domain Q&A set - 1
  - TectoMT
    - Language-independent blocks only
    - + Spanish blocks (new + adapted)
    - + domain-specific dictionary

Evaluation

- Test-sets:
  - 1,000 IT-domain Q&A set - 2
  - WMT11 newswire test-set
- Results
  - Moses outperforms the TectoMT systems
  - BLEU increases as TectoMT customisation increases
  - en->es scores higher than es->en in accordance with the development effort
  - Systems score better for the IT set

	English-Spanish		Spanish-English	
	IT	WMT11	IT	WMT11
Moses	28.12	26.91	31.92	25.24
TectoMT – language independent blocks	12.40	8.38	12.34	8.17
TectoMT – + Spanish blocks	23.62	13.92	14.67	8.50
TectoMT – + domain dictionary	26.40	13.25	15.82	8.23

Conclusions

- Development of an entry-level deep-syntax system for the English-Spanish pair
  - Reuse of English analysis and synthesis modules
  - Integration of ixa-pipes for Spanish
  - Crafting of blocks for Spanish
  - Traininig of statistical models for transfer
  - Training of morphological models for Spanish synthesis
- Available at: https://github.com/ufal/treex
- BLEU scores still behind Moses (but close for En-Es on the IT domain!)
  - Flexible customization options
  - Further customization and tuning has potential for improvement

Thank you