# Rich morpho-syntactic descriptors for factored machine translation with highly inflected languages as target

Alexandru Ceausu

Centre for Next Generation Localisation, Dublin City University

Workshop on Machine Translation and Morphologically-rich Languages
University of Haifa, 23-27 January, 2011

# Motivation

- The baseline phrase-based translation approach has limited success on translating between languages with very different syntax and morphology
- The translation is especially difficult when the direction is from a language with fixed word structure to a highly inflected language
- There are two main points to improve on:
  - morphological translation equivalence
  - long range reordering

# Introduction

- Factored translation models (Koehn şi Hoang, 2007) allow the integration of the morpho-syntactic information into the translation model.
- We present a factored translation system that uses lemma translations and morpho-syntactic correspondences to generate the target word-form.
- The experiments were carried out on a small parallel corpus (English-Bulgarian, English-Greek, English-Romanian and English-Slovenian). We show how the system scales-up to an automatically annotated corpus of 1.5 million sentence pairs (English-Romanian).
- Also, we present a method for rich morpho-syntactic annotation of highly inflected languages, considering the fact that encoding the morpho-lexical properties of the word-forms requires a large set of morpho-syntactic description codes (MSD).

# Related work

- Morphological splitting and stemming
- Supertags
  - CCG (Combinatorial Categorial Grammar) tags (Birch et al; Haque et al)
  - Syntax-to-morphology mapping (Yeniterzi & Oflazer; Avramidis & Koehn)
- Tree-based models

# Tagging with morpho-syntactic description codes (MSD)

# Morpho-syntactic description (MSD) codes

The notation format has the following main characteristics:

- attributes are marked by positions;
- values are represented by a single character;
- the character at position 0 encodes part-of-speech;
- each character at position 1, 2, …n encodes the value of one attribute (person, gender, number, etc.);
- if an attribute does not apply, it is marked with the hyphen ('-').

Ncmsrn **frate** (brother)
Ncmson **frate** (of/to a_brother)
Ncmsry **fratele** (the_brother)
Ncmsoy **fratelui**
(the_brother’s / to the_brother)

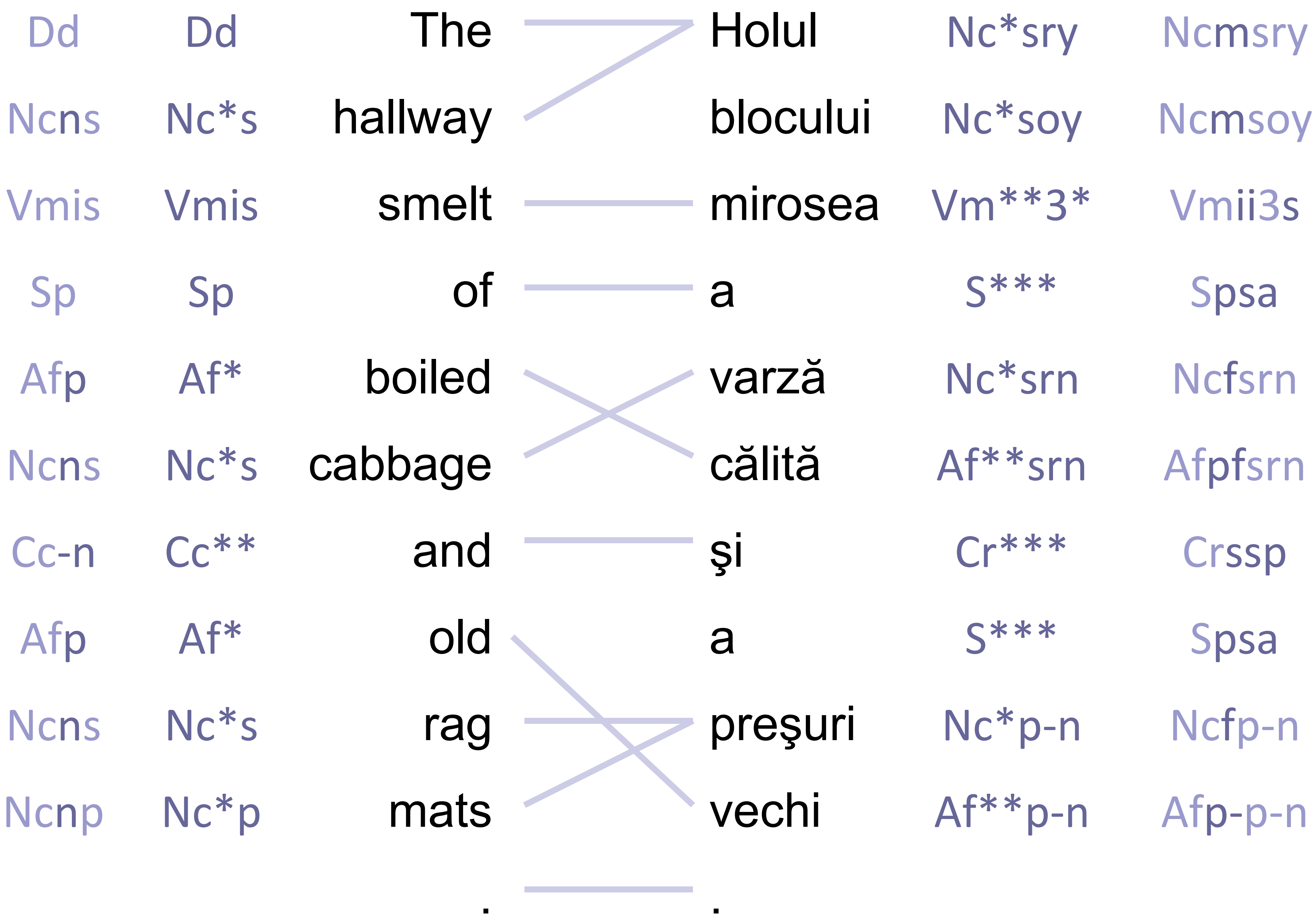
Ncmprn **fraţi** (brothers)
Ncmpon **fraţi** (of/to some brothers)
Ncmpry **fraţii** (the_brothers)
Ncmpoy **fraţilor**
(the_brothers’ / to the_brothers)

# Multext-East tag-sets

- The size of the EAGLES compliant tag-sets build within the MULTEXT-EAST initiative (Erjavec, 2004):
  - English – 133
  - Romanian – 614
  - Hungarian – 618
  - Estonian – 639
  - Czech – 1428
  - Slovene – 2083

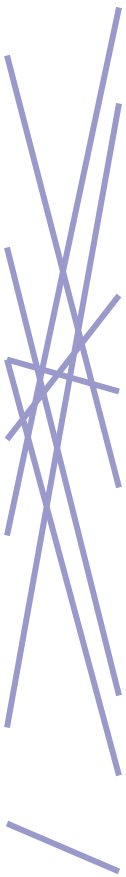
# Tiered tagging

- Tiered tagging (Tufiş, 1999) is a two-stage technique for morpho-syntactical annotation.
  - Tiered tagging uses an intermediary tag-set of a smaller size on the basis of which a language model (LM) is built. This LM serves for the first level of tagging.
  - Then, a second phase replaces the tags from the small tag-set with contextually the most probable tags from the large tag-set.



| | | | | | |
|---|---|---|---|---|---|
| Dd | Dd | The | Holul | Nc*sry | Ncmsry |
| Ncns | Nc*s | hallway | blocului | Nc*soy | Ncmsoy |
| Vmis | Vmis | smelt | mirosea | Vm**3* | Vmii3s |
| Sp | Sp | of | a | S*** | Spsa |
| Afp | Af* | boiled | varză | Nc*srn | Ncfsrn |
| Ncns | Nc*s | cabbage | călită | Af**srn | Afpfsrn |
| Cc-n | Cc** | and | şi | Cr*** | Crssp |
| Afp | Af* | old | a | S*** | Spsa |
| Ncns | Nc*s | rag | preşuri | Nc*p-n | Ncfp-n |
| Ncnp | Nc*p | mats | vechi | Af**p-n | Afp-p-n |
| | | . | . | | |

| | | |
|---|---|---|
| Afp | Af* | Vivid |
| , | , | , |
| Afp | Af* | beautiful |
| Ncnp | Nc*p | hallucinations |
| Vmis | Vmis | flashed |
| Sp | Sp | through |
| Ds3---sm | Ds****s* | his |
| Ncns | Nc*s | mind |
| | | . |



| | | | |
|---|---|---|---|
| Prin | S*** | Spsa | |
| minte | Nc*srn | Ncfsrn | |
| îi | Pp**sd********* | Pp3-sd--------w | |
| trecuseră | Vm**3* | Vmil3p | |
| nişte | Di* | Di3 | |
| vii | Af**p*n | Afp-p-n | **Afpfprn** |
| şi | Cr*** | Crssp | |
| frumoase | Af**p*n | Afpfp-n | **Afpfprn** |
| halucinaţii | Nc*p*n | Ncfprn | |
| . | | | |

# Reduced tag-set – POS tags

- The lexicon contains the words annotated with the MSD tags. For Romanian, this lexicon contains almost 1,200,000 entries.
- The reduced tag-set for Romanian consists of 92 tags plus punctuation marks.
- The reduced tag-set is derived from the MSD tag-set by repeated generalisations (leaving out some attributes from the original tag-set specification).

# Problems of the rule and lexicon-driven tiered tagging approach

- The ambiguities from the recovering process have to be solved using some additional knowledge resource (hand-written contextual disambiguation rules).
- The successful recovering is applicable only for the words recorded in the MSD tag-set lexicon.

# Tag-set conversion

- previous tags
- previous MSD features*
- suffix (1-4 characters)
- upper case (lower, all, initial)
- abbreviation (true, false)
- multiple-word expression (true, false)
- has number (true, false)
- hyphen position (none, start, middle, end)
- prefix (1-2 characters)
- word length (in characters)
- end of sentence punctuation mark

| | | |
|---|---|---|
| Dd | Dd | The |
| Ncns | Nc*s | hallway |
| Vmis | Vmis | smelt |
| Sp | Sp | of |
| Afp | Af* | boiled |
| Ncns | Nc*s | cabbage |
| Cc-n | Cc** | and |
| Afp | Af* | old |
| Ncns | Nc*s | rag |
| Ncnp | Nc*p | mats |
| . | . | . |

# Factored translation experiments

Workshop on Machine Translation and Morphologically-rich Languages
University of Haifa, 23-27 January, 2011

# SEE-ERA.net corpus

- 1204 documents from the JRC-Acquis corpus
- 60,389 translation units

| Language | No. of tokens | Avg no. of tokens/sentence |
|---|---|---|
| Bulgarian | 1,436,925 | 23.79 |
| English | 1,466,912 | 24.29 |
| Greek | 1,469,642 | 24.33 |
| Romanian | 1,422,995 | 23.56 |
| Slovene | 1,271,011 | 21.04 |

# SEE-ERA.net corpus

```
<tu id="60389"><seg lang="en"><s id="32005L0004.n.26.1.en"><w lemma="do"
ana="Vmps">Done</w><w lemma="at" ana="Sp">at</w><w lemma="Brussels"
ana="Np">Brussels</w><c>,</c><w lemma="19" ana="Mc">19</w><w lemma="January"
ana="Ncns">January</w><w lemma="2005"
ana="Mc">2005</w><c>.</c></s></seg></tu>
```

```
<tu id="60389"><seg lang="ro"><s id="32005L0004.n.26.1.ro"><w lemma="adopta"
ana="Vmp--sf">Adoptată</w><w lemma="la" ana="Spsa">la</w><w lemma="Bruxelles"
ana="Np">Bruxelles</w><c>,</c><w lemma="19" ana="Mc">19</w><w
lemma="ianuarie" ana="Ncms-n">ianuarie</w><w lemma="2005"
ana="Mc">2005</w><c>.</c></s></seg></tu>
```

```
<tu id="60389"><seg lang="sl"><s id="32005L0004.n.25.1.sl"><w lemma="v"
ana="Sl">V</w><w lemma="Bruselj" ana="Npmsl">Bruslju</w><c>,</c><w lemma="19."
ana="Mdo">19.</w><w lemma="januar" ana="Ncmsg">januarja</w><w lemma="2005"
ana="Mdm">2005</w></s></seg></tu>
```

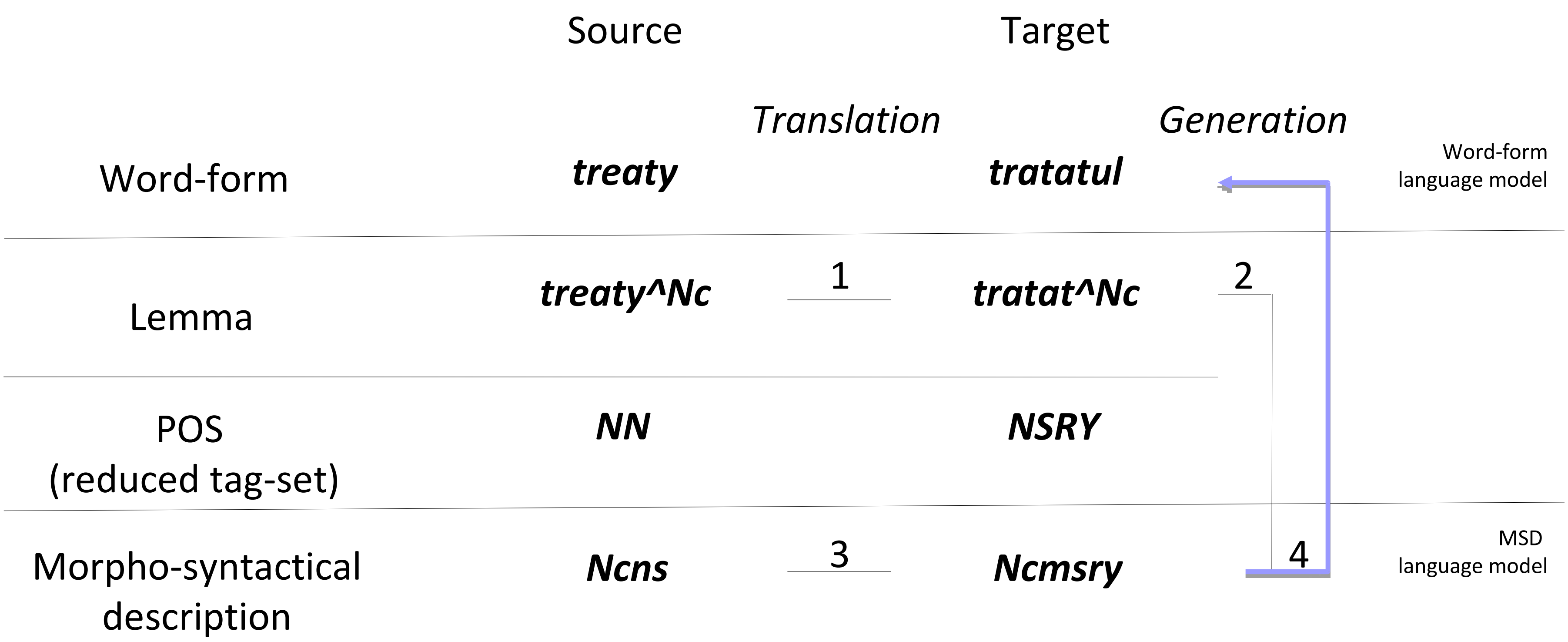
# Factored translation steps

- Translation
- Language model
- Reordering
- Generation

# Factored translation models

- Aligning and translating *lemma* could add a significant improvement especially for languages with rich morphology.
- *Part of speech affinities*. In general, the translated words tend to keep their part of speech and when this is not the case, the part-of-speech chosen is not random.
- The *re-ordering* of the target sentence words can be improved if a language model over Part-of-Speech tags is used.

# Decoding



| | Source | *Translation* | Target | *Generation* | |
|---|---|---|---|---|---|
| Word-form | ***treaty*** | | ***tratatul*** | | Word-form<br>language model |
| Lemma | ***treaty^Nc*** | 1 | ***tratat^Nc*** | 2 | |
| POS<br>(reduced tag-set) | ***NN*** | | ***NSRY*** | | |
| Morpho-syntactical<br>description | ***Ncns*** | 3 | ***Ncmsry*** | 4 | MSD<br>language model |

# Translation steps for English-Romanian

| Translation model | Generation model | Language model | Distortion model | BLEU score |
|---|---|---|---|---|
| Word-form | | Word-form | | 51.76 |
| Lemma | lemma -> word-form | Word-form | | 51.79 |
| Lemma<br>POS | lemma -> POS<br>lemma,POS -> word-form | POS<br>Word-form | | 52.31 |
| Lemma<br>MSD | lemma -> MSD<br>lemma,MSD -> word-form | MSD<br>Word-form | | 52.76 |
| Lemma<br>MSD | lemma -> MSD<br>lemma,MSD -> word-form | MSD<br>Word-form | Word-form | 46.39 |
| Lemma<br>MSD | lemma -> MSD<br>lemma,MSD -> word-form | MSD<br>Word-form | MSD | 45.77 |

Training: 58000 translation units (TU). MERT: 500 TU. Test set: 1000 TU

# Translation steps for Romanian-English

| Translation model | Generation model | Language model | Distortion model | BLEU score |
|---|---|---|---|---|
| Word-form | | Word-form | | 47.22 |
| Lemma | lemma -> wordform | Word-form | | 45.62 |
| Lemma<br>POS | lemma -> POS<br>lemma,POS -> word-form | POS<br>Word-form | | 47.37 |
| Lemma<br>MSD | lemma -> MSD<br>lemma,MSD -> word-form | MSD<br>Word-form | | 46.94 |
| Lemma<br>POS | lemma -> POS<br>lemma,POS -> word-form | POS<br>Word-form | Word-form | 51.46 |
| Lemma<br>POS | lemma -> POS<br>lemma,POS -> word-form | POS<br>Word-form | POS | 51.74 |

Training: 58000 translation units (TU). MERT: 500 TU. Test set: 1000 TU

# Evaluation - SEE-ERA.net corpus

| Direction | Baseline | Factored |
|---|---|---|
| English-Bulgarian | 38.94 | 39.60 |
| English-Romanian | 51.76 | 52.76 |
| English-Slovene | 40.73 | 42.68 |

*BLEU scores

Training: 58000 translation units (TU). MERT: 500 TU. Test set: 1000 TU

# English-Romanian 1.5 million sentence pairs corpus

| Corpus | Tokens (millions) | | Sentence pairs |
|---|---|---|---|
| | English | Romanian | |
| DGT Translation Memory | 12.5 | 12 | 621 K |
| EMEA (Opus Corpus) | 10 | 11 | 698 K |
| SE Times<br>(Opus Corpus) | 4.4 | 4.7 | 166 K |
| NAACL news | 0.8 | 0.7 | 39 K |
| Raw total | 27.7 | 28,4 | 1,525 K |
| Cleaned total | 27.3 | 27,7 | 1,495 K |

# Corpus annotation

| English | Romanian |
|---|---|
| Grounds \| ground^Nc \| NNS \| Ncnp | Motive \| motiv^Nc \| NPN \| Ncfp-n |
| of \| of^Sp \| PREP \| Sp<br>non-recognition \| recognition^Nc \| NN \| Ncns | de \| de^Sp \| S \| Spsa<br>refuz \| refuz^Nc \| NSN \| Ncms-n<br>al \| al^Ts \| TS \| Tsms<br>recunoaşterii \| recunoaştere^Nc \| NSOY \| Ncfsoy |
| for \| for^Sp \| PREP \| Sp<br>judgments \| judgment^Nc \| NNS \| Ncnp | hotărârilor_judecătoreşti \|<br>hotărâre_judecătorească^Nc \| NSRN \| Ncfsrn |
| relating \| relate^Vm \| PPRE \| Vmpp<br>to \| to^Sp \| PREP\| Sp | în \| în^Sp \| S \| Spsa<br>materia \| materie^Nc \| NSRY \| Ncfsry |
| parental_responsibility \|<br>parental_responsibility^Nc \| NN \| Ncns | răspunderii_părinteşti \|<br>răspundere_părintească^Nc \| NSOY \| Ncfsoy |

# Evaluation

- Baseline 53.82
- Factored 53.41

*BLEU scores

Training: 1.5 million translation units (TU). MERT: 1000 TU. Test set: 1000 TU

# Analysis of the results

- 200 sentences from the journalistic corpus
- Noun-phrase agreement for noun phrases with a conjunction.
- Subject – predicate agreement for predicates with verbs in indicative present

# Noun-phrase agreement

- 81 noun phrases with conjunctions
  - Baseline: 61 correct
  - Factored: 75 correct
- Example:
  - Reference: 500 items of clothing and perfume
  - Baseline: 500 de articole (Ncfp-n) de îmbrăcăminte (Ncfsrn) şi parfumurilor (Ncfpoy)
  - Factored: 500 de piese (Ncfp-n) de îmbrăcăminte (Ncfsrn) şi parfumuri (Ncfp-n)

# Subject and predicate agreement

- 123 predicates with a verb in the present tense
  - Baseline: 97 correct
  - Factored: 118 correct
- Example:
  - Reference: the military spokesman, …, said
  - Baseline: purtătorul (Ncmsry) de cuvânt al armatei, …, au (Va--3p) declarat
  - Factored: purtătorul (Ncmsry) de cuvânt al armatei, …, a (Va--3s) declarat

# Conclusions

- We found that translating lemmas and morpho-syntactical descriptors (obtained with the tiered tagging process) and generating the word-forms has better results than the baseline word-form translation model
  - better noun phrase agreement
  - better long-distance subject and predicate match in gender and number
- Lemma-based translation equivalents table produce better alignments and improves the translation accuracy.

# References

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In Proceedings of ACL-08/HLT, pages 763–770, Columbus, Ohio, June

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, pp. 1535 - 1538, ELRA, Paris

Rejwanul Haque, Sudip Kumar Naskar, Yanjun Ma & Andy Way. 2009. Using Supertags as Source Language Context in SMT. In Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT-09), May 14-15, 2009, Barcelona, Spain

Philipp Koehn, and Hieu Hoang. 2007. Factored Translation Models. In Proceedings of the 2007 Joint  Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868–876, Prague, June 2007

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318

Ralph Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, 2006, pp.2142-2147

Dan Tufiş, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, and Cvetana Krstev. 2008. Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Marko Tadić, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.) Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008), pp. 145-152, Dubrovnik, Croatia, September 25-28

Reyyan Yeniterzi and Kemal Oflazer. 2010. Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 454–464, Uppsala, Sweden, 11-16 July 2010

# Acknowledgments

- PLuTO Project (ICT-PSP-250430) - European Union’s ICT Policy Support Programme / Competitiveness and Innovation Framework Programme
- STAR (IDEI 742/19.01.2009) – CNCSIS Romania

# Thank you!