# FST Trimming: Ending Dictionary Redundancy in Apertium

Matthew Marting<sup>1</sup> Kevin Brubeck Unhammer<sup>2</sup>

<sup>1</sup>St. David's School Raleigh, NC.

<sup>2</sup>Kaldera språkteknologi Stavanger, Norway unhammer+apertium@mm.st

27th May 2014

#### Matthew Marting, Kevin Brubeck Unhammer

Introduction and background

FST's in the Apertium pipeline The Problem: Redundant data

nplementation of

Preprocessing the bilingual dictionary

fictionary Moving uninflected lemma

Intersection

lt-trim in use

Ending Dictionary
Redundancy

Conclusion

FST's in the Apertium pipeline The Problem: Redundant data

A Solution: Intersection

Implementation of

Preprocessing the bilingua dictionary

refixing the bilingual ictionary

arts

rsection

-trim in use

nding Dictiona

edundancy

onclusion

Acknowledgements

Introduction and background

Implementation of lt-trim

**Ending Dictionary Redundancy** 

Conclusion

## Matthew Marting, Kevin Brubeck

## Introduction and background

FST's in the Apertium pipeline The Problem: Redundant data

A Solution:

mplementation of

Preprocessing the bilingua

Prefixing the bilingual

Moving uninflected lemma

Intersection

t-trim in use

nding Dictions

Conclusion

- Apertium: Free/Open Source, Rule-based Machine Translation platform
- Uses Ittoolbox Finite State Transducers for:
  - morph. analysis: 'fishes' to
    fish<n><pl>/fish<vblex><pres>
  - ▶ lex. transfer: fish<n><pl> to fisk<n><m><pl><DD>
  - morph. generation: fisk<n><m><pl><def> to 'fiskane'

Introduction and

#### FST's in the Apertium pipeline

A Solution

mplementation of

\_\_ C E I M

ictionary

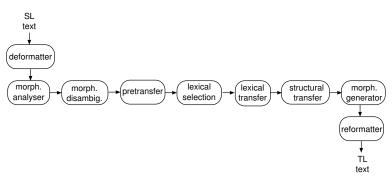
tionary

Intersection

lt-trim in uso

Ending Dictionar

Conclusion



The Problem: Redunda

Implementation of lt-trim

dictionary

Prefixing the bilingual

tionary wing uninflected lemma

rts ersection

itersection

Inding Dictionary Redundancy

onclusion

Acknowledgements

## Ittoolbox FST's support a variety of multiwords

An Ittoolbox "lexical unit" is one token, and can be:

- simple non-multi-words: 'fish'
- simple space-separated words: 'hairy frogfish' as a single token
- multiwords with inner inflection: 'takes out',
  analysed as take<vblex><pri><p3><sg># out,
  converted to take# out<vblex><pri><p3><sg> before
  lexical transfer

background

#### FST's in the Apertium pipeline

The Problem: Redund

A Solution: Int

mplementation of

Preprocessing the bilingua dictionary

lictionary

rts

ersection

t-trim in use

Ending Dictionary

Conclusion

```
pioined multiwords: 'they'll';
analysed as single token
prpers<prn><subj><p3><mf><pl>+will<vaux><inf>,
then split into two tokens
prpers<prn><subj><p3><mf><pl> and
will<vaux><inf> before lexical transfer
```

Introduction and background

#### FST's in the Apertium pipeline

The Problem: Redund

Implementation of

## lt-trim

dictionary

ictionary foving uninflected lemma

ırts

ntersection

t-trim in use

Ending Dictionary Redundancy

Conclusion

```
combinations of the 3 multiword types: 'creure-ho que',
analysed as single token
creure<vblex><inf>+ho<prn><enc><p3><nt># que,
then moved and split into two tokens
creure# que<vblex><inf> and
ho<prn><enc><p3><nt> before lexical transfer
```

## The Problem: Redundant data

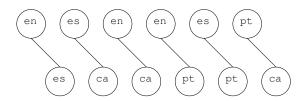


Figure: Current number of monodixes with pairs of four languages

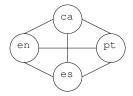


Figure: Ideal number of monodixes with four languages

It-trim

## Matthew Marting, Kevin Brubeck

ntroduction and ackground

FST's in the Apertium pipeline

#### The Problem: Redundant data

A Solution: Interse

#### npiementation of t=trim

Preprocessing the bilingual dictionary

ctionary

rts

## lt-trim in use

t-trim in use

# Redundancy

#### Conclusion

Introduction and background

FST's in the Apertium pipeline
The Problem: Bedundant data

#### A Solution: Intersection

Implementation

lt-trim

dictionary

ctionary

arts

tersection

it-trim in use

Ending Dictionary Redundancy

onclusion

Introduction and background

FST's in the Apertium pipeline The Problem: Redundant data

A Solution

#### Implementation of

lt-trim

Preprocessing the bilingu dictionary

Prefixing the bilingual dictionary

Moving uninflected lemm parts

ntersection

t-trim in use

Ending Dictionary Redundancy

Conclusion

# Preprocessing the bilingual dictionary

It-trim

#### Matthew Marting, Kevin Brubeck Unhammer

Introduction and background

FST's in the Apertium pipeline
The Problem: Redundant data

-----

lt-trim

#### Preprocessing the bilingual dictionary

dictionary

Moving uninflected lemm

ntersection

t-trim in use

Ending Dictionary Redundancy

Conclusion

# Prefixing the bilingual dictionary



#### Matthew Marting, Kevin Brubeck Unhammer

Introduction and background

FST's in the Apertium pipeline
The Problem: Bedundant data

A Solution

Implementation

Preprocessing the bilingua

Prefixing the bilingual

dictionary

ntersection

lt\_tviv in un

it-trim in use

Ending Dictionary Redundancy

Conclusion

## Moving uninflected lemma parts

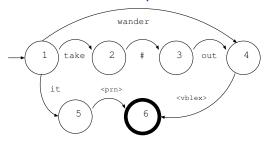
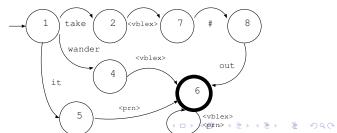


Figure: Input bilingual FST (letter transitions compressed to single arcs)



It-trim

#### Matthew Marting, Kevin Brubeck Unhammer

Introduction and background

FST's in the Apertium pipeline The Problem: Redundant data

Implementation of

Preprocessing the bilingual dictionary

ctionary

#### Moving uninflected lemma parts

ntersection

lt-trim in use

Ending Dictionary Redundancy

onclusion

## Intersection

#### It-trim

#### Matthew Marting, Kevin Brubeck Unhammer

background



#### Intersection

#### <n> b <pr>> а The Problem 6

Figure: Input monolingual FST

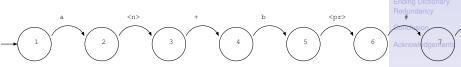


Figure: Trimmed monolingual FST

Introduction and background

FST's in the Apertium pipeline The Problem: Redundant data

A Solution

Implementation o

Preprocessing the bilingua

Prefixing the bilingual

Moving uninflected lemm parts

ntersection

 $lt-trim\ \mbox{in}\ \mbox{use}$ 

Ending Dictionary Redundancy

onclusion

## **Ending Dictionary Redundancy**



#### Matthew Marting, Kevin Brubeck Unhammer

Introduction and background

FST's in the Apertium pipeline

A Solution:

Implementation o

Preprocessing the bilingua dictionary

Prefixing the bilingual dictionary

Moving uninflected lemm parts

ntersection

lt-trim in use

### Ending Dictionary Redundancy

Conclusion

Introduction and background

FST's in the Apertium pipeline
The Problem: Redundant data

A Solution:

mplementation o

Preprocessing the bilingual dictionary

Prefixing the bilingual

Moving uninflected lemm parts

Intersection

t-trim in use

Ending Dictionary Redundancy

#### Conclusion

Introduction and background

FST's in the Apertium pipeline

A Solution

Implementation of

Preprocessing the bilingua

Prefixing the bilingual

Moving uninflected lemma

itersection

t-trim in use

Ending Dictionary Redundancy

onclusion