#### It-trim

#### Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

The Problem: Redundant data

Implementation of

Preprocessing the bilingual

Prefixing the bilingual dictionary

parts

lt\_tvim in uso

lt-trim in use

Ending Dictionary Redundancy

Conclusion

Acknowledgements

# FST Trimming: Ending Dictionary Redundancy in Apertium

Francis Tyers<sup>0</sup> Matthew Ma

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

<sup>0</sup>UiT Norgga árktalaš universitehta Romssa, Norga

ftyers@prompsit.com

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

<sup>2</sup>Kaldera språkteknologi Stavanger, Noreg

unhammer+apertium@mm.st

27th May 2014

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

A Solution: Interse

Implementation of

Preprocessing the bilingua dictionary

Prefixing the bilingual lictionary

arts

rsection

-trim in use

ndina Diation

Inding Dictionary
Redundancy

onclusion

Acknowledgements

Introduction and background

Implementation of lt-trim

**Ending Dictionary Redundancy** 

Conclusion

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

A Solution: II

nplementation of

Preprocessing the bilinguidictionary

refixing the bilingual

Moving uninflected lemr parts

t\_trimin use

t-trim in use

Ending Dictionary
Redundancy

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'

#### FST's in the Apertium pipeline

The Broblem: Bedune

A Solution

# mplementation of

t-trim

ictionary

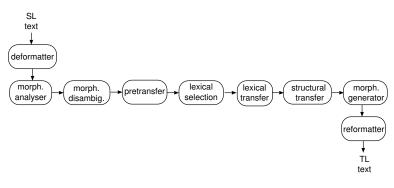
ctionary

Intersection

lt-triminus

Ending Dictionary

Conclusion



# FST's in the Apertium pipeline

A Solution: Intersection

Implementation of

Preprocessing the bilingua dictionary

refixing the bilingual ictionary

rts

tersection

t-trim in us

nding Dictionary

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

### FST's in the Apertium pipeline

The Problem: Redund

A Solution: I

mplementation of

Preprocessing the bilingua

Prefixing the bilingual dictionary

ving uninflected len ts

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
```

## FST's in the Apertium pipeline

The Problem: Redund

land and the second

# lt-trim

dictionary

Prefixing the bilingual

dictionary Moving uninflected lemma

arts

t\_twim in uso

t-trim in use

Ending Dictionary Redundancy

onclusion

```
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

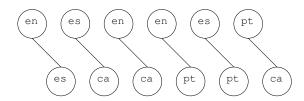


Figure: Current number of monodixes with pairs of four languages

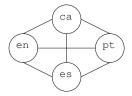


Figure: Ideal number of monodixes with four languages

It-trim

Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

FST's in the Apertium pipeline

The Problem: Redundant data

-----

mplementation of

Preprocessing the bilingual dictionary

ctionary

ts

t-trim in use

t-trim in use

Ending Dictionary
Redundancy

Conclusion

# A Solution: Intersection

It-trim

Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

A Solution: Intersection

Implementation of

lt-trim

Preprocessing the biling dictionary

Prefixing the bilingual dictionary

Moving uninflected lem parts

itersection

t-trim in use

Ending Dictionary Redundancy

onclusion

# Implementation of lt-trim

#### It-trim

### Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

A Solutio

### Implementation of

lt-trim

Preprocessing the bilingudictionary

Prefixing the bilingual dictionary

Moving uninflected lemm parts

ntersection

t-trim in use

Ending Dictionary Redundancy

onclusion

# Preprocessing the bilingual dictionary

It-trim

Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

implementation o lt-trim

Preprocessing the bilingual dictionary

dictionary

Moving uninflected lemm

tersection

-trim in use

lt-trim in use

Ending Dictionary Redundancy

Conclusion

# Prefixing the bilingual dictionary



### Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

A SUIULION

Implementation of

Preprocessing the bilingua

Prefixing the bilingual

dictionary

ntersection

t-trim in use

Ending Dictiona

onclusion

# Moving uninflected lemma parts

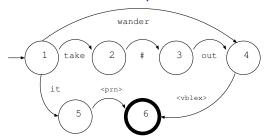
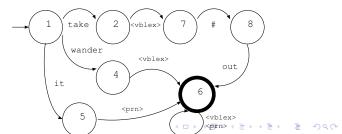


Figure: Input bilingual FST (letter transitions compressed to single arcs)



It-trim

Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

nplementation of

Preprocessing the bilingual dictionary

Moving uninflected lemma

parts

ntersection

Inding Dictionar

Redundancy

onclusion

# Intersection

а

#### It-trim

Francis Tyers, Matthew Marting, Kevin Unhammer



background background The Problem A Solution: Intersection

# Intersection



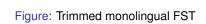


Figure: Input monolingual FST

<n>

<n>

b

<pr>>

<pr>>

6

6

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

A Solution

Implementation o

Preprocessing the bilingu

Prefixing the bilingual dictionary

Moving uninflected lemr parts

itersection

lt-trim in use

Ending Dictionary Redundancy

Conclusion

# **Ending Dictionary Redundancy**

#### It-trim

### Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

A Solution:

Implementation

Preprocessing the bilingui

Prefixing the bilingual

Moving uninflected lemm parts

ntersection

lt-trim in use

# Ending Dictionary Redundancy

Conclusion

# Conclusion

It-trim

Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

A SUIULION.

Implementation o

Preprocessing the bilingu dictionary

Prefixing the bilingual dictionary

Maxima uninfloated lamm

arts

ntersection

t-trim in use

Ending Dictionary Redundancy

#### Conclusion

# Acknowledgements

#### It-trim

### Francis Tyers, Matthew Marting, Kevin Unhammer

Introduction and background

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

A Solution:

Implementation

Preprocessing the bilings

Prefixing the bilingual dictionary

Moving uninflected lemr parts

itersection

t-trim in use

Ending Dictionary Redundancy

onclusion