LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

2014/09/01
Workshop on Finite-State
Language Resources
Sofia

Inflection transducers for simple words

Éric Laporte

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Inflection

Lemma dictionaries

Generation of inflected forms

Inflection transducers

Operators

Inflection and derivation

Inflection

collect collects collected collecting

collect is the **lemma** or canonical form

Derivation

collect collection collector collective collectivism

collect is the **base**

Dubious cases

Diminutives

Inflection in Portuguese

copo	*copinho*	lemma: *copo*	features: msD

Derivation in German

Glas	*Gläschen*	lemma: *Gläschen*	features: Nsn

Inflected-form dictionary

блестящото,блестящ.A:snd
влак,влак.N+m:s0
влака,влак.N+m:c
влака,влак.N+m:sh
влако,влак.N+m:v
влакове,влак.N+m:p0
влаковете,влак.N+m:pd
влакът,влак.N+m:sl
глава,глава.N+f:s0
главата,глава.N+f:sd
глави,глава.N+f:p0
главите,глава.N+f:pd
главо,глава.N+f:v
добра,добър.A:sf0

Source: Svetla Koeva, Cvetana Krstev

Updating an inflected-form dictionary

Evolution of the language, of the domain, of spelling, of a project

Errors

Generate a version

Each version of an inflected-form dictionary can be generated from a lemma dictionary

Variations

Variation in form

Suffixes	*give*	*gave*	*given*

Variation in grammatical features

Tense/mood	infinitive	preterit	past participle
Inflectional features			

Inflection without variation in form

	hit	*hit*	*hit*

Lemma

One of the inflected forms, chosen to represent all others in the lexical entry

collect collects collected collecting

lemma: *collect*

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Inflection

Lemma dictionaries

Generation of inflected forms

Inflection transducers

Operators

Lemma dictionary

блестящ,A3
влак,N1+m
глава,N600+f
добър,A4
индианец,N2+m
индиански,A2
индиец,N3+m
индийски,A2
индикация,N603+f
индия,N601+f+NProp
кораб,N8+m
корабче,N301+n
ладия,N603+f
лице,N300+n
лодка,N602+f
лондонски,A2
мъж,N4+m
параход,N8+m
Париж,N7+m+Nprop
плавателен,A5
съд,N1+m
фракция,N603+f
франция,N601+f+NProp
французин,N9+m+NProp
французки,A2
червен,A3
член,N5+m
човек,N6+m

Source: Cvetana Krstev

The DELAS format

влак,N1+m

lemma | part of speech | **inflectional behaviour** | other information

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Inflection

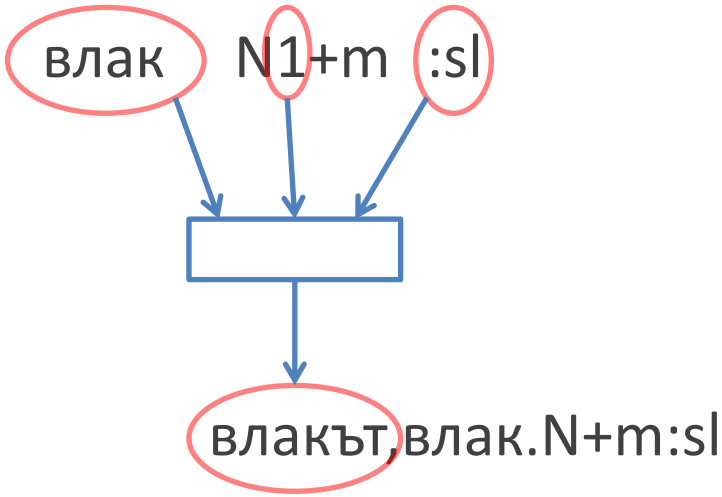
Lemma dictionaries

Generation of inflected forms

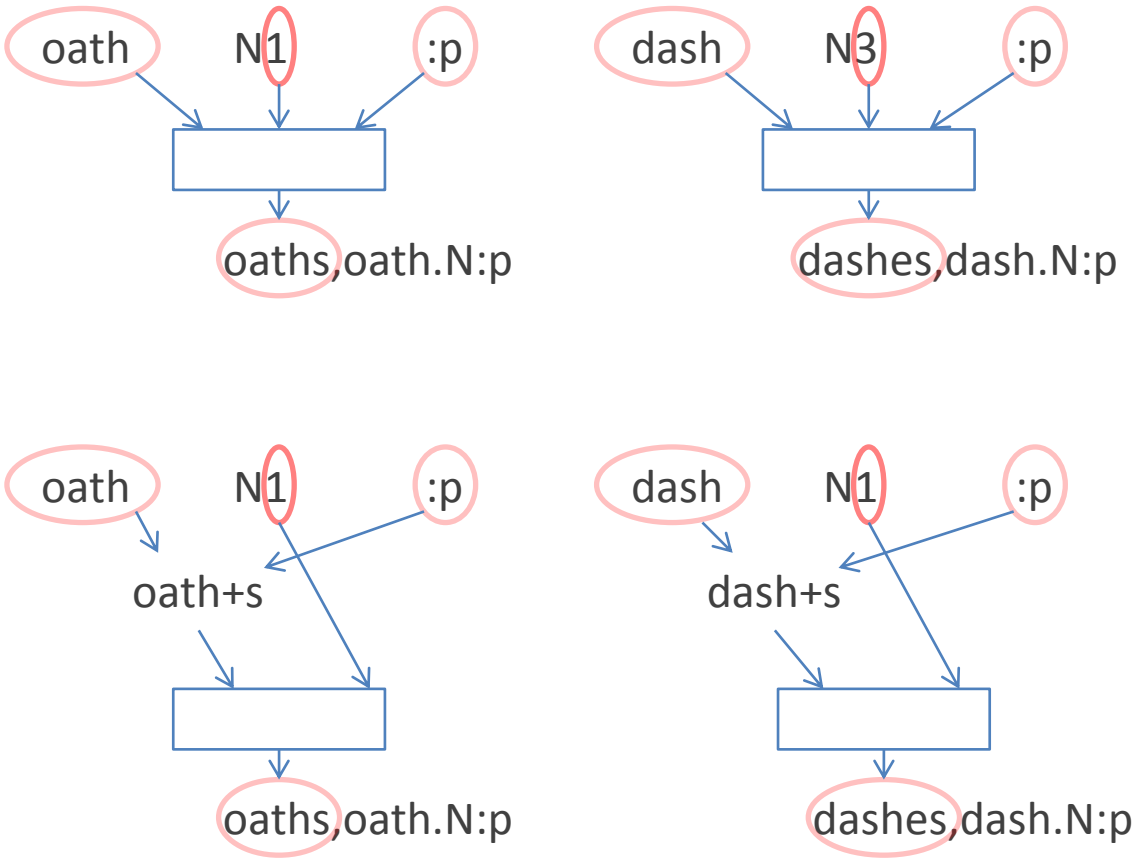
Inflection transducers

Operators

Generation of inflected forms



Two approaches



Taxonomic approach

Detailed taxonomy of inflectional behaviours

Strengths: readable transducers, updatability

Tools: Unitex-Gramlab

Context-based approach

Context-sensitive rules

Strength: Non-redundant transducers

Tools: two-level morphology

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Inflection

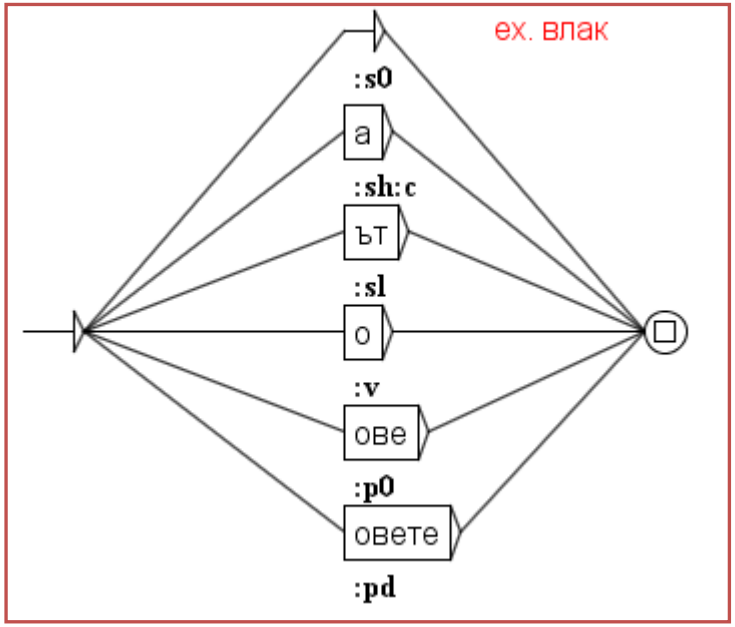
Lemma dictionaries

Generation of inflected forms

Inflection transducers

Operators

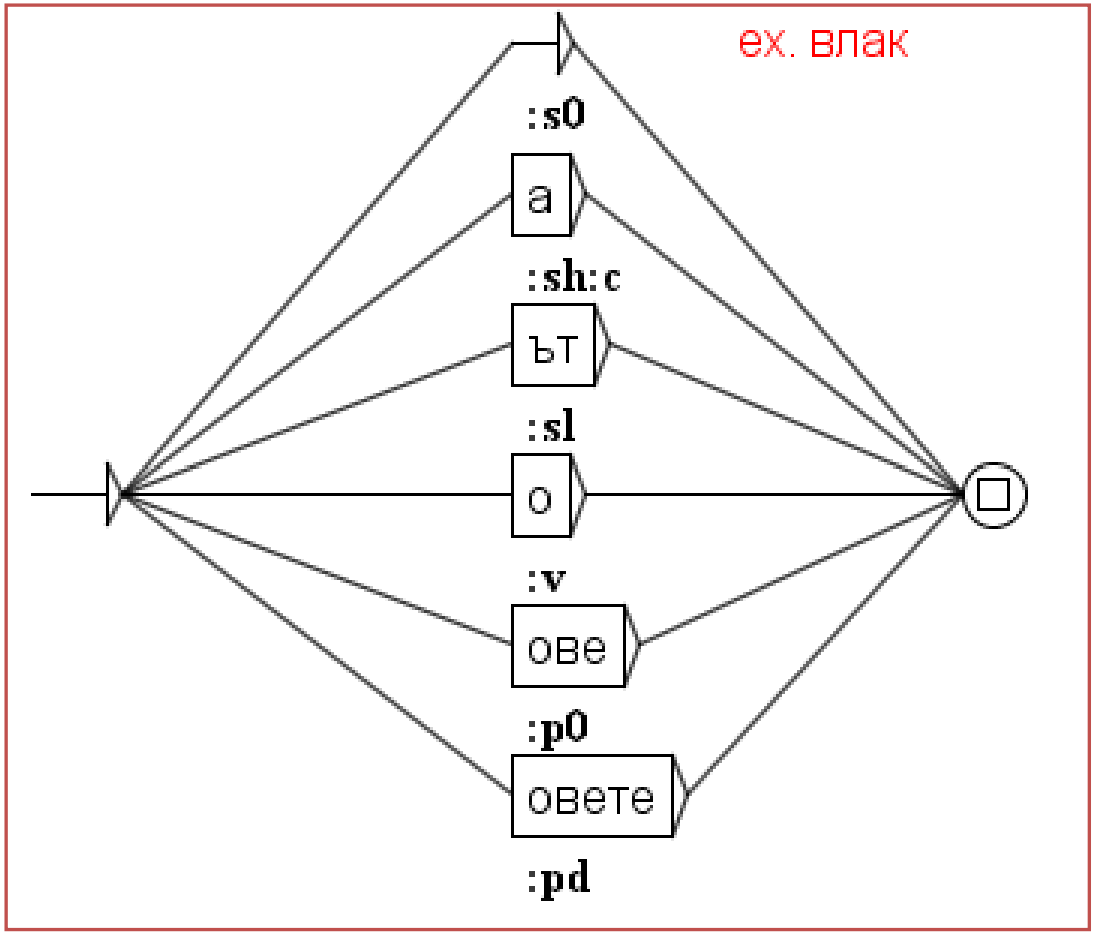
The taxonomic approach to generation of inflected forms



Inflection transducer for *влак*
Source: Cvetana Krstev

SILBERZTEIN, Max. 1998. "INTEX: An integrated FST toolbox", in Derick WOOD, Sheng YU (eds.), *Automata Implementation*, p. 185-197, Lecture Notes in Computer Science, vol. 1436. Second International Workshop on Implementing Automata (1997), Berlin/Heidelberg: Springer.

The taxonomic approach to generation of inflected forms



In the boxes

Suffixes to be appended to the lemma

Operators to edit the lemma

Below the boxes

Encoded inflectional features

Name of the transducer

N1.grf

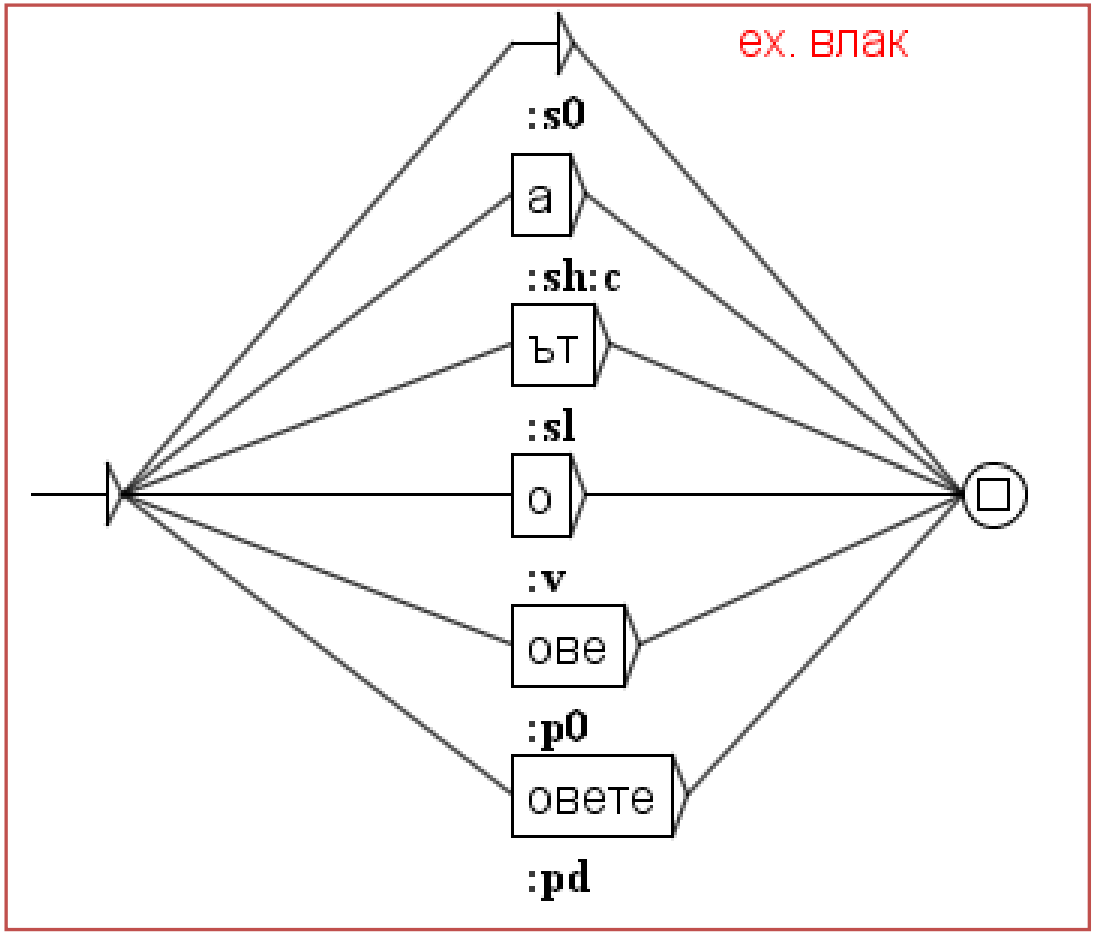
Same as the code for the inflectional behaviour

Level of generality

All nouns with inflectional behaviour modelled by this transducer

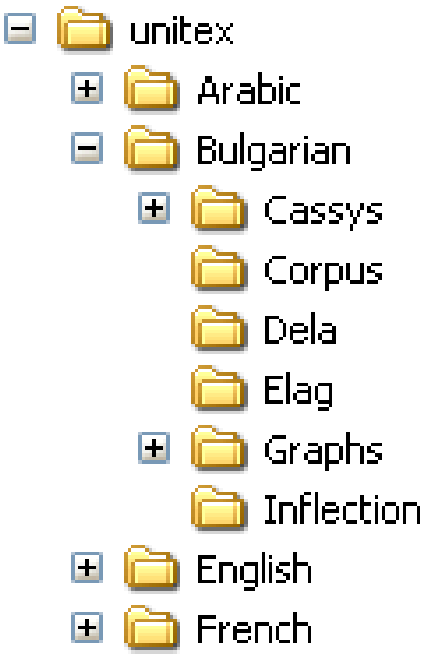
Inflection class

How to create and edit an inflection transducer with Unitex



Type in the boxes
<E>/:s0
a/:sh:c
ът/:sl
...

How to generate an inflected-form dictionary with Unitex



Save the lemma dictionary in the Dela directory of the language, with extension .dic

Save the inflection transducers in the Inflection directory

With the DELA menu of Unitex

> Check Format choose the DELA format

> Inflect choose "Allow only simple words"

> Compress into FST

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Outline

Inflection

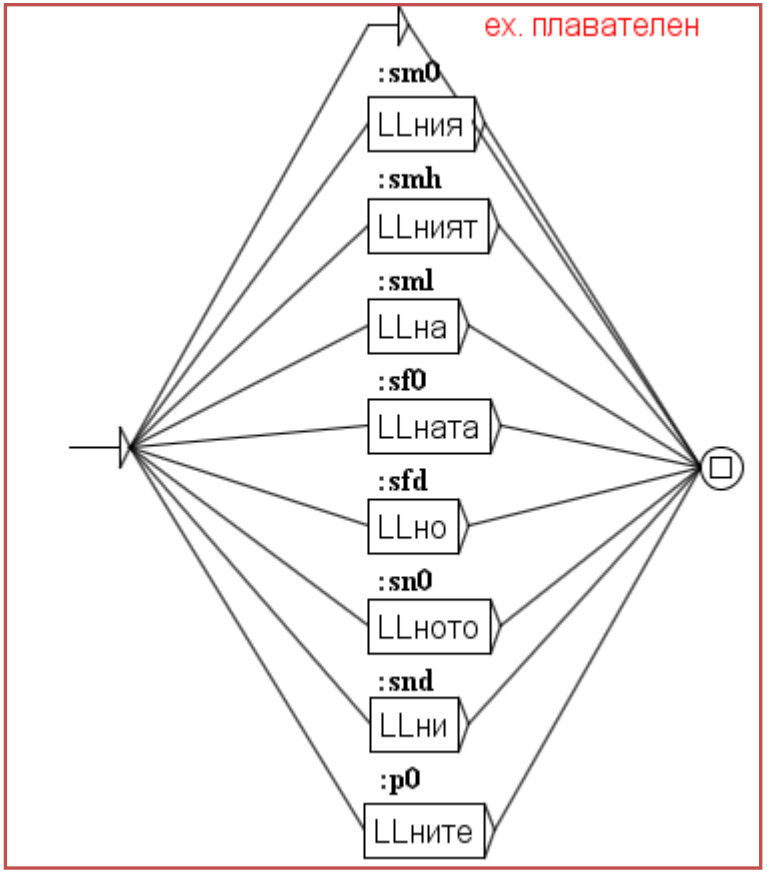
Lemma dictionaries

Generation of inflected forms

Inflection transducers

Operators

Operators to modify the lemma



Operators modify the lemma before appending a suffix

L (for *left*): delete last letter

плавателен
LLният

плавателе
LLният

плавател
LLният

плавателният
LLният

Operators to modify the lemma

Last letters	
L (for *left*)	push the last letter onto the stack
D	delete the last letter
R (for *right*)	pop one letter from the stack
C	duplicate the last letter
U	unaccent the last letter
and others	see the manual (Paumier, 2002)
First letters	
P	capitalize the first letter
W	lower-case the first letter
<I=?>	insert ? before the first letter
<X=*n*>	delete the first *n* letters
<R=?>	replace the first letter with ?

Operators to modify the lemma

плавателен — LDRият

плавателе — *н* — LDRият

плавател — *н* — LDRият

плавателн — — LDRият

плавателният — — LDRият

L (for *left*): push the last letter onto the stack

D: delete the last letter

R (for *right*): pop one letter from the stack

Any remaining letter in the stack is discarded at the end

Operators to modify the lemma

French *appeler "call"*

appeler

DDCe –

appele

DDCe –

appel

DDCe –

appell

DDCe –

appelle

DDCe –

C: duplicate the last letter

English

hit

Cing –

hitt

Cing –

hitting

Cing –

Operators to modify the lemma

Portuguese *miúdo* "tiny"

U: unaccent the last letter

miúdo	DLURinho	_
miúd	DLURinho	_
miú	DLURinho	*d*
miu	DLURinho	*d*
miud	DLURinho	_
miudinho	DLURinho	_

Operators to modify the lemma

Serbian *trom* "sluggish"

<l=?>: insert ? before the first letter

trom

<l=j><l=a><l=n>ijem

–

jtrom

<l=j><l=a><l=n>ijem

–

ajtrom

<l=j><l=a><l=n>ijem

–

najtrom

<l=j><l=a><l=n>ijem

–

najtromijem

<l=j><l=a><l=n>ijem

–

Details

Inflection transducers may have subgraphs

They may not contain lexical masks referring to information in dictionaries

The inflection tool preserves the case (upper vs. lower) of letters in lemmas and suffixes

The two approaches and updatability of transducers

Taxonomic approach

An update of a transducer does not affect inflection in other inflection classes

It is easy to control the evolution of the transducers

Context-based approach

Most rules apply to any entry

Only exceptions have conditions of application which take into account inflection classes

An update of a rule may affect the inflection of any entry

It is difficult to predict the consequences of a change

LABORATOIRE D'INFORMATIQUE
GASPARD-MONGE

Sous la co-tutelle de :
CNRS
ÉCOLE DES PONTS PARISTECH
ESIEE PARIS
UPEM • UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Thanks

CONTACT

ÉRIC LAPORTE

00 +33 (0)1 60 95 75 52

ERIC.LAPORTE@UNIV-PARIS-EST.FR