

Language Technology

Chapter 12: Words, Parts of Speech, and Morphology

Pierre Nugues

Pierre.Nugues@cs.lth.se

September 19, 2024



The Parts of Speech

The parts of speech (POS) are classes that correspond to the lexical – or word – categories

Plato made a distinction between the verb and the noun.

After him, the word categories further evolved and grew in number until Dionysus Thrax formulated and fixed them.

Aelius Donatus popularized the list of the eight parts of speech: noun, pronoun, verb, participle, conjunction, adverb, preposition, and interjection.

Grammarians have adopted these POS for most European languages although they are somewhat arbitrary

POS divide between two main classes: the open class and the closed class



Parts of Speech: Open Class Words

POS	English	French	German
Nouns	<i>name, Frank</i>	<i>nom, François</i>	<i>Name, Franz</i>
Adjectives	<i>big, good</i>	<i>grand, bon</i>	<i>groß, gut</i>
Verbs	<i>to swim</i>	<i>nager</i>	<i>schwimmen</i>
Adverbs	<i>rather, very, only</i>	<i>plutôt, très, uniquement</i>	<i>fast, nur, sehr, endlich</i>



Parts of Speech: Closed Class Words

POS	English	French	German
Determiners	<i>the, several, my</i>	<i>le, plusieurs, mon</i>	<i>der, mehrere, mein</i>
Pronouns	<i>he, she, it</i>	<i>il, elle, lui</i>	<i>er, sie, ihm</i>
Prepositions	<i>to, of</i>	<i>vers, de</i>	<i>nach, von</i>
Conjunctions	<i>and, or</i>	<i>et, ou</i>	<i>und, oder</i>
Auxiliaries and modals	<i>be, have, will, would</i>	<i>être, avoir, pouvoir</i>	<i>sein, haben, können</i>



Annotation with Parts of Speech

Sentence:

That round table might collapse

Annotation:

Words	Parts of speech	POS tags
that	Determiner	DET
round	Adjective	ADJ
table	Noun	NOUN
might	Modal verb	AUX
collapse	Verb	VERB

The automatic annotation uses predefined POS tagsets such as the Penn Treebank tagset for English



Training Sets: The CoNLL Format

The CoNLL format is a tabular format to distribute annotated texts. This format was created for evaluations carried out by the Conference in natural language learning

The CoNLL annotation has varied much across the years. We use CoNLL-U, the latest iteration.

Annotation of the Spanish sentence:

La reestructuración de los otros bancos checos se está acompañando por la reducción del personal

'The restructuring of Czech banks is accompanied by the reduction of personnel'



Example of Annotation (CoNLL-U)

La reestructuración de los otros bancos checos se está acompañando por la reducción del personal

ID	FORM	LEMMA	UPOS	FEATS
1	La	el	DET	Definite=Def Gender=Fem Number=Sing PronType=Art
2	reestructuración	reestructuración	NOUN	Gender=Fem Number=Sing
3	de	de	ADP	AdpType=Prep
4	los	el	DET	Definite=Def Gender=Masc Number=Plur PronType=Art
5	otros	otro	DET	Gender=Masc Number=Plur PronType=Ind
6	bancos	banco	NOUN	Gender=Masc Number=Plur
7	checos	checo	ADJ	Gender=Masc Number=Plur
8	se	se	PRON	Case=Acc Person=3 PrepCase=Npr PronType=Prs Reflex=Yes
9	está	estar	AUX	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin
10	acompañando	acompañar	VERB	VerbForm=Ger
11	por	por	ADP	AdpType=Prep
12	la	el	DET	Definite=Def Gender=Fem Number=Sing PronType=Art
13	reducción	reducción	NOUN	Gender=Fem Number=Sing
14	del	del	ADP	AdpType=Preppron
15	personal	personal	NOUN	Gender=Masc Number=Sing
16	.	.	PUNCT	PunctType=Peri



Part-of-Speech Annotation (CoNLL 2000)

Annotation of: *He reckons the current account deficit will narrow to only # 1.8 billion in September.* We set aside the last column for now.

He	PRP	B-NP
reckons	VBZ	B-VP
the	DT	B-NP
current	JJ	I-NP
account	NN	I-NP
deficit	NN	I-NP
will	MD	B-VP
narrow	VB	I-VP
to	TO	B-PP
only	RB	B-NP
#	#	I-NP
1.8	CD	I-NP
billion	CD	I-NP
in	IN	B-PP
September	NNP	B-NP
.	.	O



Part-of-Speech Annotation (Universal Dependencies)

Annotation of: *Do museum labels have an impact on how people look at artworks?*

ID	FORM	LEMMA	UPOS	XPOS	FEATS
1	Do	do	AUX	VBP	Mood=Ind Number=Plur Person=3 Tense=Pres VerbForm=Fin
2	museum	museum	NOUN	NN	Number=Sing
3	labels	label	NOUN	NNS	Number=Plur
4	have	have	VERB	VB	VerbForm=Inf
5	an	a	DET	DT	Definite=Ind PronType=Art
6	impact	impact	NOUN	NN	Number=Sing
7	on	on	ADP	IN	
8	how	how	SCONJ	WRB	PronType=Rel
9	people	person	NOUN	NNS	Number=Plur
10	look	look	VERB	VBP	Mood=Ind Number=Plur Person=3 Tense=Pres VerbForm=Fin
11	at	at	ADP	IN	
12	artworks	artwork	NOUN	NNS	Number=Plur
13	?	?	PUNCT	.	



Part-of-Speech Annotation (Universal Dependencies)

Annotation of: *Genom skattereformen införs individuell beskattning (särbeskattning) av arbetsinkomster.*

ID	FORM	LEMMA	UPOS	XPOS	FEATS
1	Genom	genom	ADP	PP	
2	skattereformen	skattereform	NOUN	NN UTR SIN DEF NOM	Case=Nom Definite=Def Gender=Com Number=Sing
3	införs	införa	VERB	VB PRS SFO	Mood=Ind Tense=Pres VerbForm=Fin Voice=Pass
4	individuell	individuell	ADJ	JJ POS UTR SIN IND NOM	Case=Nom Definite=Ind Degree=Pos Gender=Com Number=Sing
5	beskattning	beskattning	NOUN	NN UTR SIN IND NOM	Case=Nom Definite=Ind Gender=Com Number=Sing
6	((PUNCT	PAD	
7	särbeskattning	särbeskattning	NOUN	NN UTR SIN IND NOM	Case=Nom Definite=Ind Gender=Com Number=Sing
8))	PUNCT	PAD	
9	av	av	ADP	PP	
10	arbetsinkomster	arbetsinkomst	NOUN	NN UTR PLU IND NOM	Case=Nom Definite=Ind Gender=Com Number=Plur
11	.	.	PUNCT	MAD	



Ambiguity

Words	Possible tags	Example of use
that	Subordinating conjunction Determiner Adverb Pronoun Relative pronoun	<i>That he can swim is good</i> <i>That white table</i> <i>It is not that easy</i> <i>That is the table</i> <i>The table that collapsed</i>
round	Verb Preposition Noun Adjective Adverb	<i>Round up the usual suspects</i> <i>Turn round the corner</i> <i>A big round</i> <i>A round box</i> <i>He went round</i>
table	Noun Verb	<i>That white table</i> <i>I table that</i>
might	Noun Modal verb	<i>The might of the wind</i> <i>She might come</i>
collapse	Noun Verb	<i>The collapse of the empire</i> <i>The empire can collapse</i>



Part-of-Speech Ambiguity in Swedish

The word *som* in the *Norstedts svenska ordbok*, 1999, has three entries:

- ① *Om jag vore lika vacker som du, skulle jag vara lycklig.* (konjunktion)
- ② *Bilen som jag köpte i fjol.* (pronomen)
- ③ *Som jag har saknat dig.* (adverb)

The part-of-speech difference can be significant:

Swedish. Compare the pronunciation of *vaken*, adjective, as in *Han är aldrig vaken innan klockan sju* and *vaken*, noun, as in *Vi fiskade i vaken i sjön*

English. Compare *object* in *I object to violence*, verb, or *I could see an object*, noun.



Grammatical Features

Main parts of speech	Features (subcategories)
Adjective, noun, pronoun	Regular base comparative superlative interrogative person number case
Adverb	Regular base comparative superlative interrogative
Article, determiner, preposition	Person case number
Verb	Tense voice mood person number case



Lexicons: An Excerpt from the Oxford Advanced Learner's Dictionary

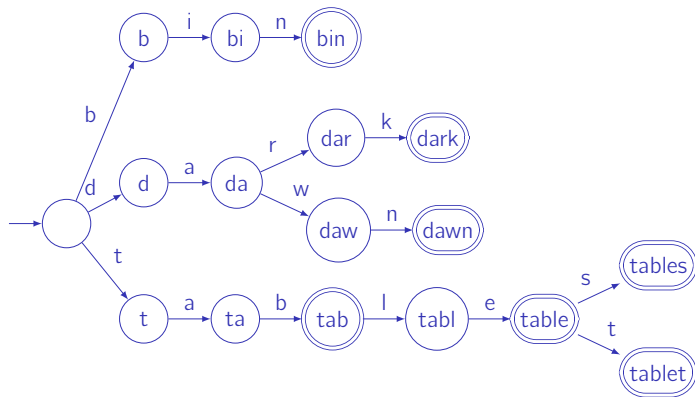
Word	Pronunciation	Syntactic tag	Syllable count or verb pattern (for verbs)
a	@	S-*	1
a	EI	Ki\$	1
a fortiori	el ,fOtl'Oral	Pu\$	5
a posteriori	el ,pOsterl'Oral	OA\$,Pu\$	6
a priori	el ,pral'Oral	OA\$, Pu\$	4
a's	Eiz	Kj\$	1
ab initio	&b I'nISl@U	Pu\$	5
abaci	'&b@sal	Kj\$	3
aback	@'b&k	Pu%	2
abacus	'&b@k@s	K7%	3
abacuses	'&b@k@slz	Kj%	4
abaft	@'bAft	Pu\$,T-\$	2
abandon	@'b&nd@n	H0%,L@%	36A,14
abandoned	@'b&nd@nd	Hc%,Hd%,OA%	36A,14
abandoning	@'b&nd@nIN	Hb%	46A,14
abandonment	@'b&nd@nm@nt	L@%	4
abandons	@'b&nd@nz	Ha%	36A,14
abase	@'bels	H2%	26B
abased	@'belst	Hc%,Hd%	26B
abatement	@'belsm@nt	L@%	3

Newer standard:

<https://tei-c.org/release/doc/tei-p5-doc/en/html/DI.html>



Letter Trees



Morphemes

	Word	Morpheme decomposition
English	<i>disentangling</i>	<i>dis+en+tangle+ing</i>
	<i>rewritten</i>	<i>re+write+en</i>
French	<i>désembrouillé</i>	<i>dé+em+brouiller+é</i>
	<i>récite</i>	<i>re+écrire+te</i>
German	<i>entwirrend</i>	<i>ent+wirren+end</i>
	<i>wiedergeschrieben</i>	<i>wieder+ge+schreiben+en</i>



Inflection

	Plural of nouns	Morpheme decomposition
English	<i>hedgehogs</i>	<i>hedgehog+s</i>
	<i>churches</i>	<i>church+es</i>
	<i>sheep</i>	<i>sheep+∅</i>
French	<i>hérissons</i>	<i>hérisson+s</i>
	<i>chevaux</i>	<i>cheval+ux</i>
German	<i>Gründe</i>	<i>Grund+(¨)e</i>
	<i>Hände</i>	<i>Hand+(¨)e</i>
	<i>Igel</i>	<i>Igel+∅</i>



Derivation

Creation of a new word

	English	French	German
Prefixes	fore see, un pleasant	pré voir, dé plaisant	vor hersehen, un angenehm
Suffixes	manage able , rigor ous	gér able , rigour eux	vorsicht ich , streit bar



Morphological Processing

Generation →

English		French		German	
<i>dog+s</i>	<i>dogs</i>	<i>chien+s</i>	<i>chiens</i>	<i>Hund+e</i>	<i>Hunde</i>
<i>work+ing</i>	<i>working</i>	<i>travailler+ant</i>	<i>travaillant</i>	<i>arbeiten+end</i>	<i>arbeitend</i>
<i>un+do</i>	<i>undo</i>	<i>dé+faire</i>	<i>défaire</i>		

← Parsing



Language Differences (Source: Xerox)

Language	# stems	# inflected forms	Lex. size (kb)
English	55,000	240,000	200–300
French	50,000	5,700,000	200–300
German	50,000	350,000 or infinite (compounding)	450
Japanese	130,000	200 suffixes	500
		20,000,000 word forms	500
Spanish	40,000	3,000,000	200–300



Two-Level Morphology

Current morphological parsers are based on the two-level model of Kimmo Koskenniemi (1983).

It links the surface form of a word – the word as it is in a text – to its lexical or underlying form – its sequence of morphemes

Surface:	disentangled
Lexical (or underlying):	dis+en+tangle+ed



Examples

Generation: Lexical to surface form →

English	<i>dis+en+tangle+ed</i>	<i>disentangled</i>
	<i>happy+er</i>	<i>happier</i>
	<i>move+ed</i>	<i>moved</i>
French	<i>dés+em+brouiller+é</i>	<i>désembrouillé</i>
	<i>dé+chanter+erons</i>	<i>déchanterons</i>
German	<i>ent+wirren+end</i>	<i>entwirrend</i>
	<i>wieder+ge+schreiben+en</i>	<i>wiedergeschrieben</i>

Parsing: ← Surface to lexical form



Aligning the Two Forms

English	dis+en+tangle+ed ⇕ ... dis0en0tangl00ed	happy+er ⇕ ... happi0er	move+ed ⇕ ... mov00ed
French	dé+chanter+erons ⇕ ... dé0chant000erons	cheval+ux ⇕ ... cheva00ux	cheviller+é ⇕ ... chevill000é
German	singen+st ⇕ ... sing00st	Grund+`e ⇕ ... Gründ00e	Igel+Ø ⇕ ... Igel00



Interpreting the Morphemes

Suffixes have a grammatical interpretation: *erons* in a French verb corresponds to verb + future + 1st person + plural

Morphological parsers can represent the lexical form as a concatenation of the stem and its features instead of the stem and the suffix.

The Xerox parser output for *disentangled* and *happier* is:

disentangle+Verb+PastBoth+123SP

happy+Adj+Comp

where +Verb denotes a verb, +PastBoth, either past tense or past participle, and +123SP any person, singular or plural; +Adj denotes an adjective and +Comp, a comparative.



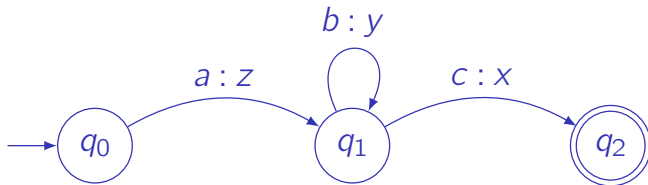
Aligning Morphemes and Features

Lexical:	d	i	s	e	n	t	a	n	g	l	e	+Verb	+PastBoth	+123sp
	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕		↕	↕
Surface:	d	i	s	e	n	t	a	n	g	l	0	0	e	d

Lexical:	h	a	p	p	y	+Adj	+Comp
	↕	↕	↕	↕	↕	↕	↕
Surface:	h	a	p	p	i	e	r



Transducers



The string *abbbc* is transduced into *zyyyx*



Mathematical Definition of a FST

- ① Q is a finite set of states.
- ② Σ is a finite set of symbol or character pairs $i : o$, where i is a symbol of the input alphabet and o of the output alphabet. As we saw, both alphabets may include epsilon transitions.
- ③ q_0 is the start state, $q_0 \in Q$.
- ④ F is the set of final states, $F \subseteq Q$.
- ⑤ δ is the transition function $Q \times \Sigma \rightarrow Q$, where $\delta(q, i, o)$ returns the state where the automaton moves when it is in state q and consumes the input symbol pair $i : o$.

The quintuple defining automaton is $Q = \{q_0, q_1, q_2\}$,

$\Sigma = \{a : z, b : y, c : x\}$,

$\delta = \{\delta(q_0, a : z) = q_1, \delta(q_1, b : y) = q_1, \delta(q_1, c : x) = q_2\}$, and $F = \{q_0\}$.



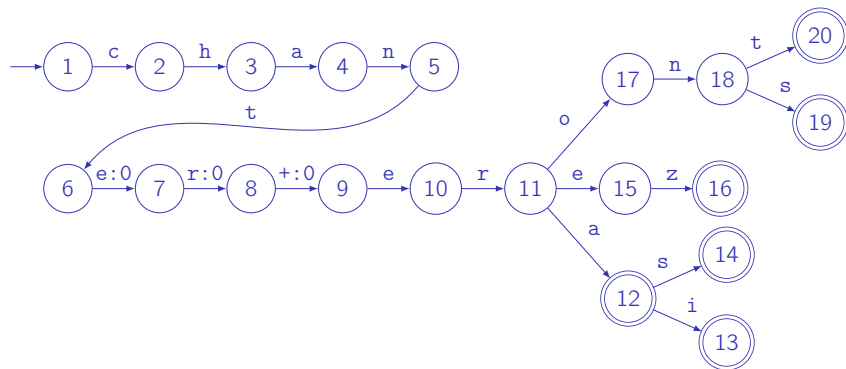
French Verb Transducers for *chanter*

Number\Person	First	Second	Third
singular	<i>chanterai</i>	<i>chanteras</i>	<i>chantera</i>
plural	<i>chanterons</i>	<i>chanterez</i>	<i>chanteront</i>

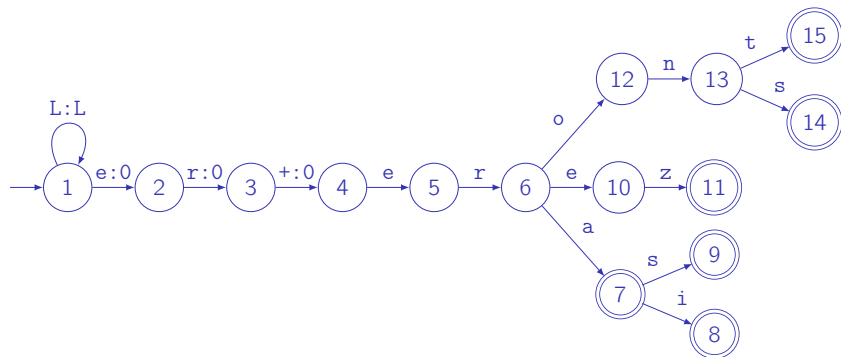
Number\Pers.	First	Second	Third
singular	chanter+erai	chanter+eras	chanter+era
	chant000erai	chant000eras	chant000era
plural	chanter+erons	chanter+erez	chanter+eront
	chant000erons	chant000erez	chant000eront



Transducer for *chanter*



French Verb Transducers: Future, 1st Group



To implement it, you can use either Prolog or OpenFst
<https://www.openfst.org/>



Romance Languages

Language	Number\Person	First	Second	Third
Italian				
	singular	<i>canterò</i>	<i>canterai</i>	<i>canterà</i>
	plural	<i>canteremo</i>	<i>canterete</i>	<i>canteranno</i>
Spanish				
	singular	<i>cantaré</i>	<i>cantarás</i>	<i>cantará</i>
	plural	<i>cantaremos</i>	<i>cantaréis</i>	<i>cantarán</i>
Portuguese				
	singular	<i>cantarei</i>	<i>cantarás</i>	<i>cantará</i>
	plural	<i>cantaremos</i>	<i>cantareis</i>	<i>cantarão</i>



Course Content: 2023

The rest of the slides in this document is not part of the course in 2023.

You can nonetheless read it if you are curious.



Ambiguity

In the transducer for future tense, there is no ambiguity: A surface form has only one lexical form with a unique final state.

This is not the case with the present tense

(je) chante 'I sing'

(il) chante 'he sings'

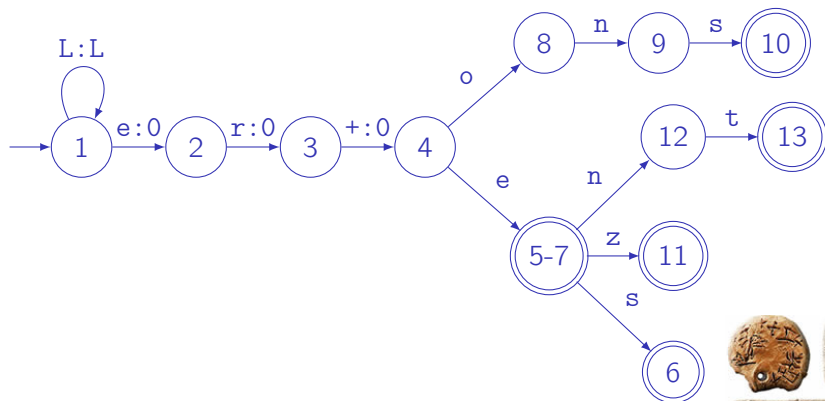
Number\Person	First	Second	Third
singular	chante	chantes	chante
plural	chantons	chantez	chantent



Transducer Ambiguity

Final states 5 and 7 are the same.

The implementation in Prolog is similar to that of the future tense. Using backtracking, the transducer can produce all the final states reflecting the morphological ambiguity.



Koskenniemi's Rules

Koskenniemi described morphology with declarative rules.

They use the left and right context and the \Rightarrow , \Leftarrow , \Leftrightarrow , or $/\Leftarrow$ operators

In English, a lexical *y* can correspond to a surface *i* as in *happier*.

It occurs when *y* is preceded by a consonant and followed by *-er*, *-ed*, or *-s*.

① $y:i \Leftarrow C:C _ _ +:0 \ e:e \ r:r$

② $y:i \Leftarrow C:C _ _ +:e \ s:s$

③ $y:i \Leftarrow C:C _ _ +:0 \ e:e \ d:d$



Two-level Rules

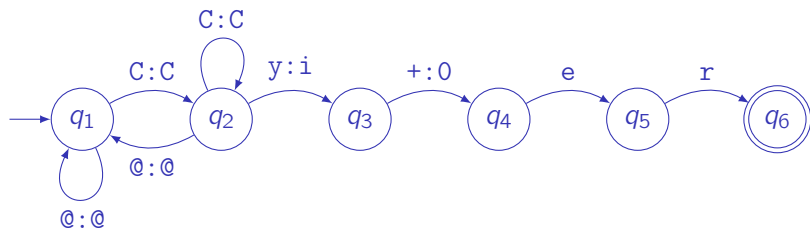
Lexical:surface transduction is described by rules.

Rules	Description
a:b \Rightarrow lc __ rc	a is transduced as b only when it has lc to the left and rc to the right
a:b \Leftarrow lc __ rc	a is always transduced as b when it has lc to the left and rc to the right
a:b \Leftrightarrow lc __ rc	a is transduced as b always and only when it has lc to the left and rc to the right
a:b $/\Leftarrow$ lc __ rc	a is never transduced as b when it has lc to the left and rc to the right



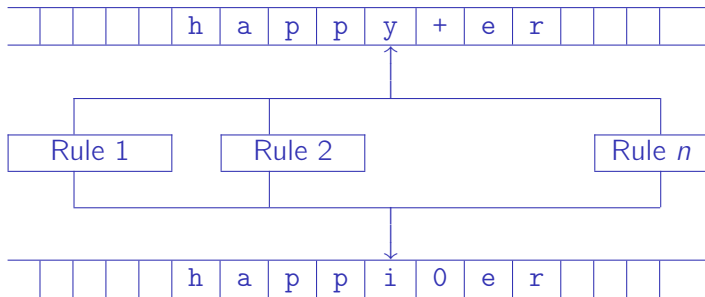
Parallel Rules

All the rules are applied in parallel (provided that their context match)



Rules and Transducers

Rules can be compiled as an equivalent transducer



Rule Intersection

The parallel transducers are then combined into a single one using the transducer intersection.



Problems with Intersection

The intersection of two finite automata defines a finite-state automaton. It is not always the case for finite-state transducers.

Kaplan and Kay (1994) demonstrated that when surface and lexical pairs have the same length – without ϵ –, the intersection is a transducer.

This property is sufficient to intersect the rules in practical applications.

In fact, transducers obtained from two-level rules are intersected by treating the ϵ symbol as an ordinary symbol (Beesley and Karttunen 2003, p. 55).



Xerox

Originally, rules were compiled by hand.

However, it can quickly become intractable especially when it comes to managing conflicting rules or when rule contexts interfere with transduced symbols.

To solve it, we can use a compiler that creates transducers automatically from two-level rules.

The Xerox's XFST is an example of it. It is a publicly available tool and to date the only serious implementation of a morphological rule compiler.



Morphology of French Verbs

We used the stem and a set of suffixes for French regular verbs.

French irregular verbs are notoriously more complex.

Chanod (1994) gives an example of decomposition into simple rules.

Infinitive	courir	dormir	battre	peindre	écrire
First person sing.	cours <u>s</u>	dors	bats	peins	écris
Second person sing.	cours <u>s</u>	dors	bats	peins	écris
Third person sing.	court <u>t</u>	dort	bat	peint	écrit
First person pl.	cour <u>ons</u>	dormons	battons	peignons	écrivons
Second person pl.	cour <u>ez</u>	dormez	battez	peignez	écrivez
Third person pl.	cour <u>ent</u>	dorment	battent	peignent	écrivent



French Morphology

Lexical form: stem	dormir	+IndP	+SG	+P1
	↕		↕	
Intermediate form: inflection	dorm	+IndP	+SG	+P1
	↕		↕	
Intermediate form: deletion of <i>m</i> followed by <i>s</i>	dorm		s	
	↕		↕	
Surface form:	dor		s	

From *peindre* to *peins*

n:0 ⇔ g __ [s|t]



Composition and Intersection

