Derivational paradigms: pushing the analogy

Olivier Bonami[1] & Jana Strnadová[2]

[1]Université Paris Diderot
[2]Google, Inc.

Paradigms in Word Formation @ SLE, August 2016

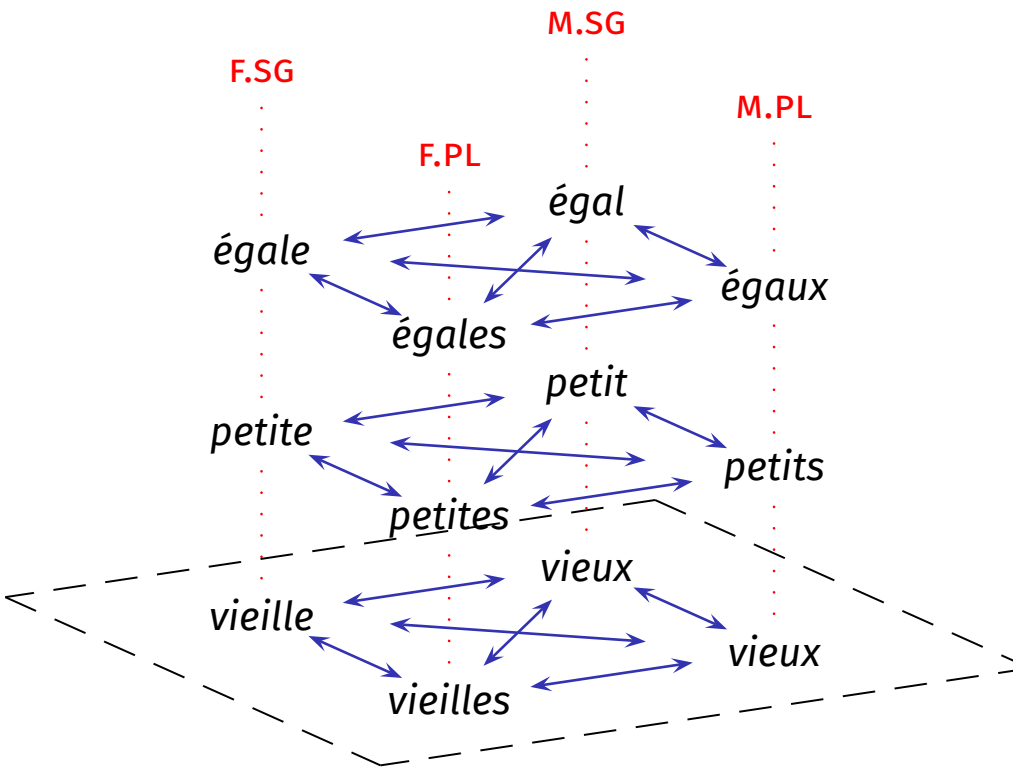
Introduction

- Two ways of using the notion of paradigm in word formation:
  1. Focus on paradigmatic relations between lexemes rather than syntagmatic relations between words and word parts.
     (see e.g. van Marle, 1984; Becker, 1993; Booij, 2010)
  2. Extend to derivational (sub)families analytic techniques developed for the study of inflectional paradigms.
     (see e.g. Matthews, 1972; Stump, 2001; Ackerman and Malouf, 2013)
- Here we take the second approach.
- We show that:
  1. Collections of structured derivational (sub)families exhibit key properties shared by inflection systems.
  2. Quantitative techniques designed for the study of inflectional paradigms can be applied fruitfully to derivational (sub)families.
- We exemplify with data from French.

Some definitions

- Morphological subfamily
  Set of words that are morphologically related.
  - ⇒ sets of words, not lexemes
  - ⇒ not necessarily exhaustive sets
- Paradigmatic system
  Collection of morphological subfamilies structured by the same system of oppositions of content (cf. Štekauer 2014) charaterized by morphosyntacic property sets.
- Paradigm
  One member of a paradigmatic system.

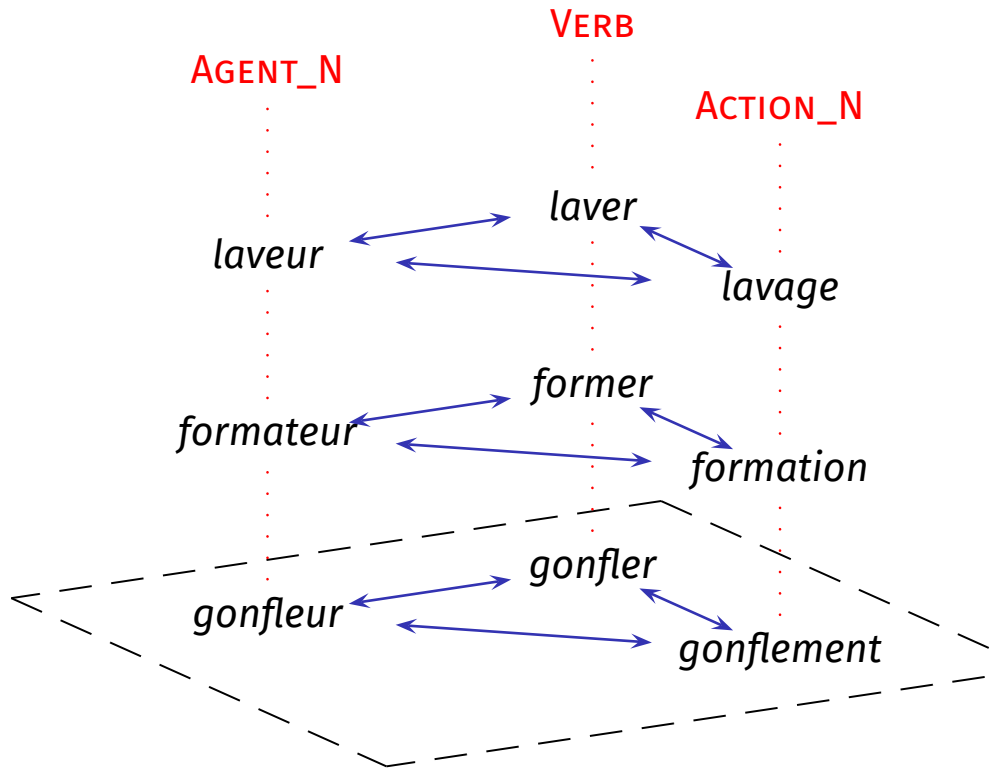
Inflectional example:



Some definitions

- **Morphological subfamily**
  Set of words that are morphologically related.
  - ⇒ sets of words, not lexemes
  - ⇒ not necessarily exhaustive sets
- **Paradigmatic system**
  Collection of morphological subfamilies structured by the same system of oppositions of content (cf. Štekauer 2014).
- **Paradigm**
  One member of a paradigmatic system.

Derivational example:



Remarks

Note that:

- We do not define paradigmatic systems as exhaustive, neither vertically nor horizontally.
- Our definition of paradigmatic systems does not allow for gaps (defectivity) or synonymy within a paradigm (overabundance).
  - Overabundance and defectiveness are just ignored
  - So are partial productivity and semantic drift

$\Rightarrow$ We focus on those cases where inflection and derivation are maximally similar, and avoid discussing how dissimilar they are in other situations.

- Paradigms are structured sets of words, but a paradigm may contain multiple inflected forms of multiple lexemes.
  - For simplicity, when dealing with derivation, we focus on systems with only one form per lexeme.

Fruitful analogies

Differential exponence

- In a paradigmatic system, the same contrasts may be encoded in different ways for different paradigms.
- This is true both for inflectionally and derivationally-related words.

	NOM.SG	GEN.PL	
(a)	hrad	hradů	‘castle’
(b)	žena	žen	‘woman’
(c)	táta	tátů	‘dad’
(d)	stavení	stavení	‘building’

Partial inflectional paradigms of a few Czech nouns

	AREA		INHABITANT	
(a)	France	‘France’	Français	‘French’
(b)	Russie	‘Russia’	Russe	‘Russian’
(c)	Albanie	‘Albania’	Albanais	‘Albanian’
(d)	Corse	‘Corsica’	Corse	‘Corsican’

Partial paradigms of French toponyms and related demonyms

Orthogonality of content and marking

- In a paradigmatic system, the formally unmarked cell (if any) need not be the same for all paradigms.
- This is true both for inflectionally and derivationally-related words.

	NOM.SG	GEN.PL	
(a)	hrad	hradů	‘castle’
(b)	žena	žen	‘woman’
(c)	táta	tátů	‘dad’
(d)	stavení	stavení	‘building’

Partial inflectional paradigms of a few Czech nouns

	AREA		INHABITANT	
(a)	France	‘France’	Français	‘French’
(b)	Russie	‘Russia’	Russe	‘Russian’
(c)	Albanie	‘Albania’	Albanais	‘Albanian’
(d)	Corse	‘Corsica’	Corse	‘Corsican’

Partial paradigms of French toponyms and related demonyms

Heteroclisis

- In a paradigmatic system, some paradigms may use an exponence strategy that is a hybrid of two others.
- This is true both for inflectionally and derivationally-related words.

	NOM.SG	GEN.PL	
(a)	hrad	hradů	‘castle’
(b)	žena	žen	‘woman’
(c)	táta	tátů	‘dad’
(d)	stavení	stavení	‘building’

Partial inflectional paradigms of a few Czech nouns

	AREA		INHABITANT	
(a)	France	‘France’	Français	‘French’
(b)	Russie	‘Russia’	Russe	‘Russian’
(c)	Albanie	‘Albania’	Albanais	‘Albanian’
(d)	Corse	‘Corsica’	Corse	‘Corsican’

Partial paradigms of French toponyms and related demonyms

Syncretism

- In a paradigmatic system, some paradigms may fail to contrast formally words that contrast in content.
- This is true both for inflectionally and derivationally-related words.

	NOM.SG	GEN.PL	
(a)	hrad	hradů	‘castle’
(b)	žena	žen	‘woman’
(c)	táta	tátů	‘dad’
(d)	stavení	stavení	‘building’

Partial inflectional paradigms of a few Czech nouns

	AREA		INHABITANT	
(a)	France	‘France’	Français	‘French’
(b)	Russie	‘Russia’	Russe	‘Russian’
(c)	Albanie	‘Albania’	Albanais	‘Albanian’
(d)	Corse	‘Corsica’	Corse	‘Corsican’

Partial paradigms of French toponyms and related demonyms

Distribution of syncretism I

- In inflection, different paradigms give rise to different patterns of syncretism.

NOM	GEN	DAT	ACC	LOC	INS	
host	hosta	hostovi, hostu	hosta	hostovi, hostu	hostem	‘guest’
lingvista	lingvisty	lingvistovi	lingvistu	lingvistovi	lingvistou	‘linguist’
most	mostu	mostu	most	mostu, mostě	mostem	‘bridge’
věta	věty	větě	větu	větě	větou	‘sentence’
kost	kosti	kosti	kost	kosti	kostí	‘bone’
město	města	městu	město	městě, městu	městem	‘city’

Distribution of syncretism II

- Within derivational paradigms too, different paradigms give rise to different patterns of syncretism.

institution		member	of institution	of member
académie	'academy'	académicien	académique	académique
sénat	'senate'	sénateur	sénatorial	sénatorial
ministère	'ministry'	ministre	ministériel	ministériel
école	'school'	écolier	scolaire	écolier
prison	'prison'	prisonnier	carcéral	prisonnier
lycée	'high school'	lycéen	lycéen	lycéen
parlement	'parliament'	parlementaire	parlementaire	parlementaire

The quantitative study

Looking ahead

- For now, we have shown how analytic concepts designed for inflection can fruitfully be applied to derivational paradigms.
- We now show how information-theoretic measures of paradigm structure inform us on relations within derivational families.
  - We specifically use the tools of Bonami and Beniamine (inpress).
  - This elaborates on much previous work; see e.g. Ackerman et al. (2009); Ackerman and Malouf (2013); Blevins (in press); Bonami and Boyé (2014); Bonami and Luís (2014); Sims (2015)
- The plan:
  1. Definition and illustration of implicative entropy
  2. Characterization of our dataset
  3. Results

The quantitative study

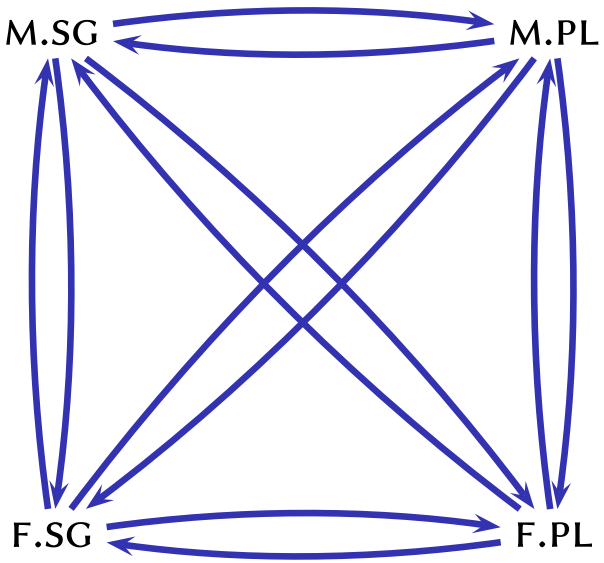
1. Implicative entropy

Predictivity in inflectional paradigms

When a speaker knows only one form of a lexeme, how hard is it to predict the others?

(Ackerman et al. (2009)'s Paradigm Cell Filling Problem)

Consider French adjectives:



- F.SG⇒F.PL is trivial
- M.SG⇒M.PL is easy but not trivial, see /lokal/∼/loko/ vs. /banal/∼/banal/
- F.SG⇒M.SG is harder, see /lɛd/∼/lɛ/ vs. /ʁɛd/∼/ʁɛd/
- M.SG⇒F.SG is hardest, see /gɛ/∼/gɛ/ vs. /lɛ/∼/lɛd/ vs. /njɛ/∼/njɛz/ vs. ...

Implicative entropy, by example

Lexeme	M.SG	M.PL	alternation M.SG ∼ M.PL	M.SG shape $\text{M.SG}_{\text{M.SG}\sim\text{M.PL}}$
LOYAL	lwajal	lwajo	*X*al ∼ *X*o	ends in -al
BANAL	banal	banal	*X* ∼ *X*	
CALME	kalm	kalm	*X* ∼ *X*	does not end in -al
POLI	poli	poli	*X* ∼ *X*	

Data sample: French masculine adjectives

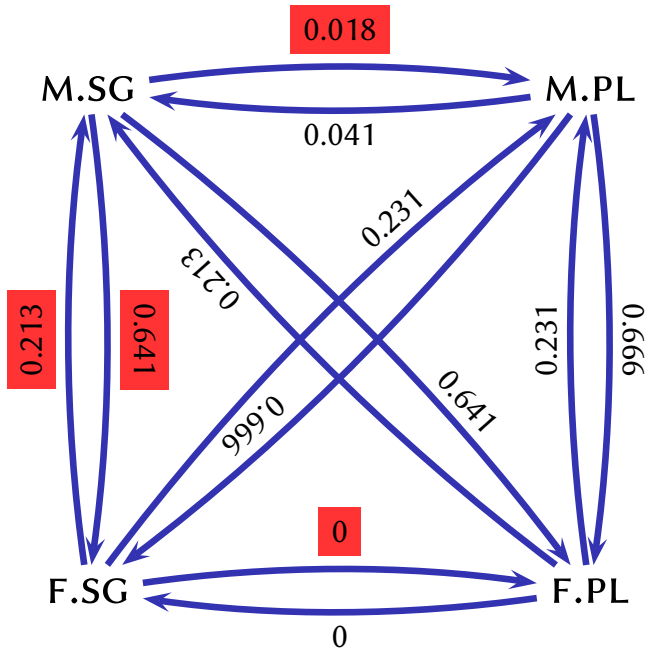
- Group lexemes by type of alternation: M.SG ∼ M.PL
- Group M.SG by shape, on the basis of which alternations these shapes are compatible with: $\text{M.SG}_{\text{M.SG}\sim\text{M.PL}}$
- The implicative entropy from M.SG to M.PL is the conditional entropy of patterns of alternation given input cell.

$$H(\text{M.SG} \Rightarrow \text{M.PL}) = H(\text{M.SG} \sim \text{M.PL} \mid \text{M.SG}_{\text{M.SG}\sim\text{M.PL}})$$

- In our toy example, $H(\text{M.SG} \Rightarrow \text{M.PL}) = 0.5$ bit
- In *Flexique* (Bonami et al., 2014), $H(\text{M.SG} \Rightarrow \text{M.PL}) = 0.017$ bit

Differential opacity

- Some paradigm cells are good predictors, others are good predictees



- What counts as a “hard case” depends on predictor and predictee.
  - M.SG→M.PL is trivial except where M.PL ends in *-al*.
  - M.PL→M.SG is trivial except where M.PL ends in *-o*.
  - M.SG→F.SG is hardest if M.SG ends in a vowel
  - etc.

Joint predictiveness

- Bonami and Beniamine (inpress) on Romance conjugation: on average, knowing multiple forms of the same lexeme makes the PCFP a lot easier.
- For French adjectives:

1 predictor	0.2966
2 predictors	0.1443
3 predictors	0.0044

- This provides a strong argument for paradigms as first class citizens of the morphological universe: there is useful knowledge on the system that can only be attained by attending to (sub)paradigms.

The quantitative study

2. The dataset

The dataset I

- We use data from Démonette (Hathout and Namer, 2014), a database of 20,493 derivational relations between 22,570 French lexemes.

…	abandonner	@	abandon	@ACT	…
…	abandonner	@	abandonneur	@AGM	…
…	abandon	@AGT	abandonneur	@AGM	…
…	abandonner	@	abandonnement	@ACT	…
…	…	…	…		

- From Démonette we tabulate 5,414 paradigms for triples (Verb, Action noun, Masculine agent noun)

@	@ACT	@AGM
abaisser	abaissement	abaisseur
abandonner	abandon;abandonnement	abandonneur;abandonnateur
abattre	abattement;abattage	abatteur
affamer		affammeur
	agriculture	agriculteur
…	…	…

The dataset II

- Since we want to deal neither with overabundance nor with defectivity:
  1. We drop all paradigms with an unfilled cell.
  2. In cases of overabundant cells, if one cell-mate makes up $\frac{2}{3}$ or more of the distribution, we drop the other cell-mates; otherwise, we drop the whole paradigm.

@	@ACT	@AGM
abaisser	abaissement	abaisseur
abandonner	abandon~~;abandonnement~~	abandonneur~~;abandonnateur~~
~~abattre~~	~~abattement;abattage~~	~~abatteur~~
~~affamer~~		~~affammeur~~
	~~agriculture~~	~~agriculteur~~
…	…	…

⇒ 1,331 remaining canonical paradigms.

The dataset III

- To assess predictibility on the basis of phonological forms, we use transcription from the GLÀFF, a lexicon derived from French Wiktionary (Hathout et al., 2014)

@	@ACT	@AGM
a.bɛ.se	a.bɛ.smɑ̃;a.bɛs.mɑ̃	a.be.sœʁ
a.bɑ̃.dɔ.ne	a.bɑ̃.dɔ̃	a.bɑ̃.dɔ.nœʁ
…	…	…

⇒ 913 paradigms for which all transcriptions are available.

The quantitative study

3. Results

Differential opacity

	Verb	Action_N	Agent_N
Verb	—	1.115	0.709
Action_N	0.101	—	0.269
Agent_N	0.264	1.114	—

Unary implicative entropy
for (Verb, Action_N, Agent_N) triples

Differential opacity

	Verb	Action_N	Agent_N
Verb	—	1.115	0.709
Action_N	0.101	—	0.269
Agent_N	0.264	1.114	—

Unary implicative entropy
for (Verb, Action_N, Agent_N) triples

Verb	Action_N	Agent_N
laver	lavage	laveur
'wash'	'washing'	'washer'
contrôler	contrôle	contrôleur
'control'	'control'	'controller'
corriger	correction	correcteur
'correct'	'correction'	'corrector'
former	formation	formateur
'train'	'training'	'trainer'
écrire	écriture	scripteur
'write'	'writing'	'writer'
gonfler	gonflement	gonfleur
'inflate'	'inflating'	'inflater'

Sample triples

- Action nouns are hardest to predict, because of the diversity of marking strategies (*-age*, *-ment*, *-ion*, *-ure*, conversion, etc.)

Differential opacity

	Verb	Action_N	Agent_N
Verb	—	1.115	0.709
Action_N	0.101	—	0.269
Agent_N	0.264	1.114	—

Unary implicative entropy
for (Verb, Action_N, Agent_N) triples

Verb	Action_N	Agent_N
laver	lavage	laveur
'wash'	'washing'	'washer'
contrôler	contrôle	contrôleur
'control'	'control'	'controller'
corriger	correction	correcteur
'correct'	'correction'	'corrector'
former	formation	formateur
'train'	'training'	'trainer'
écrire	écriture	scripteur
'write'	'writing'	'writer'
gonfler	gonflement	gonfleur
'inflate'	'inflating'	'inflater'

Sample triples

- Verbs are easiest to predict: the only challenging cases are stem suppletion and non-first conjugation.

Differential opacity

	Verb	Action_N	Agent_N
Verb	—	1.115	0.709
Action_N	0.101	—	0.269
Agent_N	0.264	1.114	—

Unary implicative entropy
for (Verb, Action_N, Agent_N) triples

Verb	Action_N	Agent_N
laver	lavage	laveur
'wash'	'washing'	'washer'
contrôler	contrôle	contrôleur
'control'	'control'	'controller'
corriger	correction	correcteur
'correct'	'correction'	'corrector'
former	formation	formateur
'train'	'training'	'trainer'
écrire	écriture	scripteur
'write'	'writing'	'writer'
gonfler	gonflement	gonfleur
'inflate'	'inflating'	'inflater'

Sample triples

- Action nouns are good predictors of agent nouns, since they almost always use the same stem.

Differential opacity

	Verb	Action_N	Agent_N
Verb	—	1.115	0.709
Action_N	0.101	—	0.269
Agent_N	0.264	1.114	—

Unary implicative entropy
for (Verb, Action_N, Agent_N) triples

Verb	Action_N	Agent_N
laver	lavage	laveur
'wash'	'washing'	'washer'
contrôler	contrôle	contrôleur
'control'	'control'	'controller'
corriger	correction	correcteur
'correct'	'correction'	'corrector'
former	formation	formateur
'train'	'training'	'trainer'
écrire	écriture	scripteur
'write'	'writing'	'writer'
gonfler	gonflement	gonfleur
'inflate'	'inflating'	'inflater'

Sample triples

- On the other hand, verbs are not so good predictors of agent nouns, because, even in the absence of suppletion, one has to guess whether the *-at-* augment should be used.

Joint predictiveness I

- Predicting from two members of a morphological family is a lot easier than predicting from just one.

1 predictor	0.595
2 predictors	0.196

Average implicative entropy

Joint predictiveness II

- In particular, predicting the form of verbs from knowledge of the two nouns is trivial.

Predictors	Predicted	Entropy
Verb, Action_N	Agent_N	0.138
Verb, Agent_N	Action_N	0.444
Agent_N, Action_N	Verb	0.006

  - All the remaining uncertainty is caused by a handful of *-ionner* verbs (Lignon and Namer, 2010).

(Action_N , Agent_N)	⇒ Verb
(percussion , percuteur)	⇒ percuter
(inspection , inspecteur)	⇒ inspecter
(perquisition , perquisiteur)	⇒ perquisitionner
(fonction , foncteur)	⇒ fonctionner

Sample triples

Conclusions

- In this talk, we applied analytic tools originally conceived for inflection to derivational families.
  1. We confirmed that inflectional and derivational families have the same kind of paradigmatic structure.
  2. We uncovered new generalizations on predictibility within derivational families.
- While we decided to set aside differences between inflection and derivation, this has no bearing on our results.
  - ⇒ the benefits of paradigmatic analysis are available whether one takes inflection and word-formation to be disjoint or undistinguished.
- Next step: pursue the extensibility of the notion of overabundance to concurrent derivatives.
  - *original* vs. *originel* ‘original’
  - *mortel* ‘mortal‘ vs. *mortuaire* ‘mortuary’
  - etc.

References

Ackerman, F., Blevins, J. P., and Malouf, R. (2009). 'Parts and wholes: implicative patterns in inflectional paradigms'. In J. P. Blevins and J. Blevins (eds.), *Analogy in Grammar*. Oxford: Oxford University Press, 54–82.

Ackerman, F. and Malouf, R. (2013). 'Morphological organization: the low conditional entropy conjecture'. *Language*, 89:429–464.

Becker, T. (1993). 'Back-formation, cross-formation, and 'bracketing paradoxes' in paradigmatic morphology'. In G. Booij and J. van Marle (eds.), *Yearbook of Morphology 1993*. Dordrecht: Kluwer, 1–25.

Blevins, J. P. (in press). *Word and Paradigm Morphology*. Oxford: Oxford University Press.

Bonami, O. and Beniamine, S. (inpress). 'Joint predictiveness in inflectional paradigms'. *Word Structure*, 9.

Bonami, O. and Boyé, G. (2014). 'De formes en thèmes'. In F. Villoing, S. Leroy, and S. David (eds.), *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Presses Universitaires de Paris Ouest, 17–45.

Bonami, O., Caron, G., and Plancq, C. (2014). 'Construction d'un lexique flexionnel phonétisé libre du français'. In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, and S. Prévost (eds.), *Actes du quatrième Congrès Mondial de Linguistique Française*. 2583–2596.

Bonami, O. and Luís, A. R. (2014). 'Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative'. In J.-L. Léonard (ed.), *Morphologie flexionnelle et dialectologie romane. Typologie(s) et modélisation(s).*, no. 22 in Mémoires de la Société de Linguistique de Paris. Leuven: Peeters, 111–151.

Booij, G. (2010). *Construction morphology*. Oxford: Oxford University Press.

Hathout, N. and Namer, F. (2014). 'Démonette, a French derivational morpho-semantic network'. *Linguistic Issues in Language Technology*, 11:125–168.

Hathout, N., Sajous, F., and Calderone, B. (2014). 'GLÀFF, a large versatile French lexicon'. In *Proceedings of LREC 2014*.

Lignon, S. and Namer, F. (2010). 'Comment conversionner les v-ion ? ou la construction de v-ionnerverbe par conversion'. In *Actes du 2eme Congrès Mondial de Linguistique Française*. 1009–1028.

Matthews, P. H. (1972). *Inflectional Morphology. A Theoretical Study Based on Aspects of Latin Verb Conjugation*. Cambridge: Cambridge University Press.

Sims, A. (2015). *Inflectional defectiveness*. Cambridge: Cambridge University Press.

Stump, G. T. (2001). *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge: Cambridge University Press.

van Marle, J. (1984). *On the Paradigmatic Dimension of Morphological Creativity*. Dordrecht: Foris.

Štekauer, P. (2014). 'Derivational paradigms'. In R. Lieber and P. Štekauer (eds.), *The Oxford Handbook of Derivational Morphology*. Oxford: Oxford University Press, 354–369.

Directional syncretism I

	I		II		III		IV	V
	M/F	M/F	NEU	M/F	NEU	M/F	NEU	M/F
NOM	aqua	dominus	donum	homo	nomen	gradus	cornu	res
ACC	aquam	dominum	donum	hominem	nomen	gradum	cornu	rem
GEN	aquae	domini	doni	hominis	nominis	gradus	cornus	rei
DAT	aquae	domino	dono	homini	nomini	gradui	cornui	rei
ABL	aqua	domino	dono	homine	nomine	gradu	cornu	re
	water	*master*	*gift*	*man*	*name*	*step*	*horn*	*thing*

Singular declension of Latin nouns