DBnary: de la boîte à chaussure aux données liées ouvertes

Gilles Sérasset

GETALP-LIG, Université Grenoble Alpes, France

09/12/2014

Première partie I

Principes fondateurs de mon travail

1. Prendre les dictionnaires comme ils sont !
2. Ne pas se contenter d'un multilinguisme linguo-centré
3. Voir d'un même œil corpus et dictionnaire

Prendre les dictionnaires comme ils sont !

- Nombreuses structures de données utilisées dans les dictionnaires
  - automates, structures de traits...
  - ...un aspect présent dans ma thèse, mais certainement pas assez poussée dans la suite.
- S'est traduit dans la plateforme "Jibiki"
  - Qui prend en charge des dictionnaires tels qu'ils sont (en XML)...
  - ...et calque une structure "standard" (CDM) sur la structure originale.
  - utilisée dans le projet Papillon, mais aussi pour d'autres projets (LexALP, GDEF, DILAF, ...)
  - maintenant pris en charge par Mathieu Mangeot.

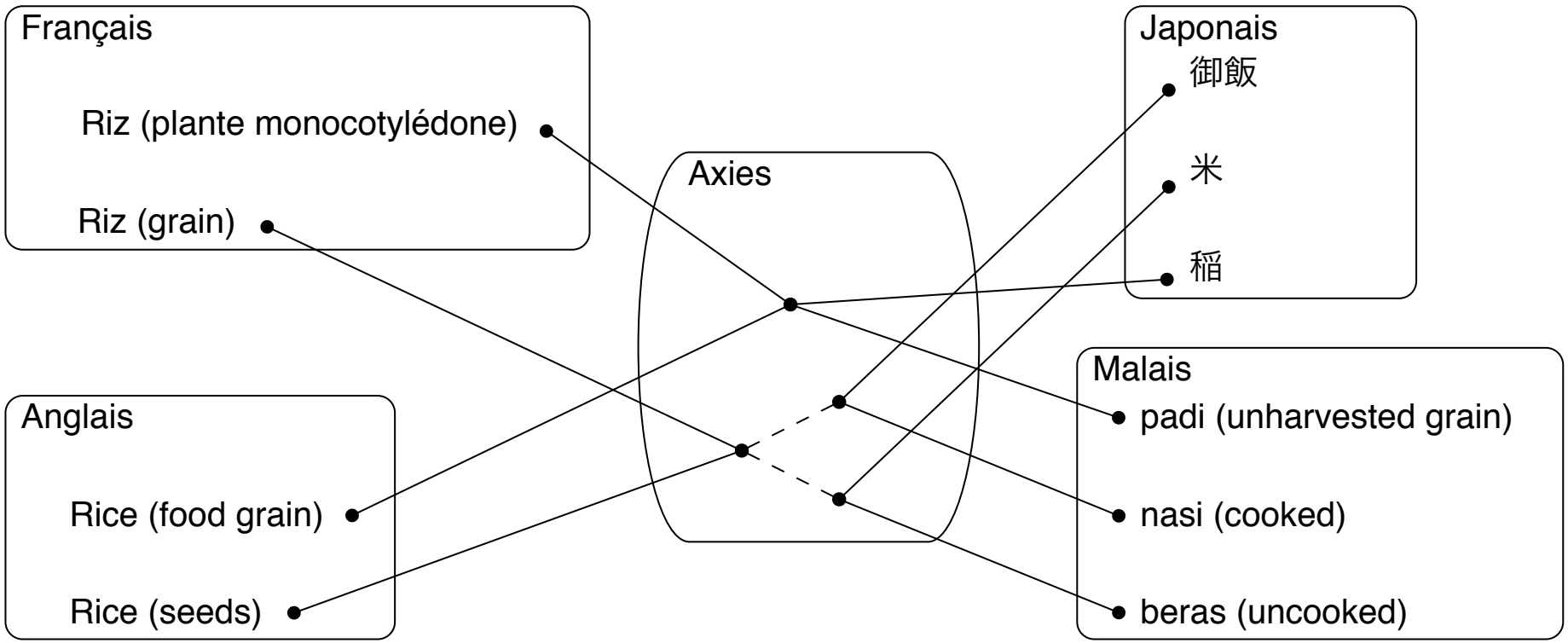
1. Prendre les dictionnaires comme ils sont !
2. Ne pas se contenter d'un multilinguisme linguo-centré
3. Voir d'un même œil corpus et dictionnaire

Ne pas se contenter d'un multilinguisme "linguo-centré"

- Aucune langue naturelle ne doit se prévaloir du rôle de pivot
  - ni l'anglais, ni le français, ni l'esperanto, ni le klingon...
- Notion d'acception interlingue (Axies)
  - Utilisée dans le projet Papillon...
  - ...ainsi que dans le projet LexALP (terminologie juridique multilingue).

Papillon : Architecture des données

Macrostructure : An acception based multilingual lexical database



Architecture of the Data

A microstructure inspired by Mel'čuk's ECD and Polguère's DICO

regretter ,

v.tr.

sentiment LA personne X ~ SON action Y

GOVERNMENT PATTERN

X = I Y = II

1 . N
1 . N
2 . de V-inf

LEXICAL FUNCTIONS

QSyn : se repentir

S0 : regret#1

Able2 : (*Que l'on peut R.*) regrettable

Magn : (*Intensément*) beaucoup

Y étant grave, Magn : amèrement , cruellement ; _se mordre les doigts_

EXAMPLES

1 . *C'est une décision qu'il va regretter cruellement.*

2 . *Il ne regrette pas d'avoir investi 4 000 F dans ce nouveau programme.*

1. Prendre les dictionnaires comme ils sont !
2. Ne pas se contenter d'un multilinguisme linguo-centré
3. Voir d'un même œil corpus et dictionnaire

Voir d'un même œil corpus et dictionnaire

- beaucoup de travaux d'extraction d'information lexicales à partir de corpus...
  - observation de co-occurences, différentes définitions de voisinages...
  - mais peu de travaux utilisent un lexique existant...
  - si les méthodes de traitement du corpus sont maîtrisées, l'utilisation d'un dictionnaire est toujours un bricolage, ad hoc...
- les dictionnaires sont de plus en plus vus comme des graphes
  - ex : les "Lexical Systems" d'A. Polguère...
- Pourquoi ne pas voir le corpus comme un graphe ?
  - Différentes topologies donnent différentes notions de voisinage...
  - Des opérations sur les graphes
    - ★ graphe d'occurence → graphe de lemmes
    - ★ fusion d'un graphe de corpus avec un graphe de dictionnaire)...
  - Thèse de Vincent Archer

Deuxième partie II

Motivations de DBnary

Un petit bilan

- Approche par acceptions interlingues (Axies)
  - toujours d'actualité, mais grosse difficulté de mise en œuvre
  - pas d'autre proposition, un besoin toujours ressenti dans les standards actuels (e.g. : ontolex)
  - des efforts en gestation (cf. dernière conf invitée de Piek Vossen à Reyjkjavik)

Un petit bilan

- Approche par acceptions interlingues (Axies)
  - toujours d'actualité, mais grosse difficulté de mise en œuvre
  - pas d'autre proposition, un besoin toujours ressenti dans les standards actuels (e.g. : ontolex)
  - des efforts en gestation (cf. dernière conf invitée de Piek Vossen à Reyjkjavik)
- Construction collaborative de bases lexicales (Papillon)
  - n'a jamais vraiment marché...
  - du fait du manque d'animation de communauté...
  - et parce que le dictionnaire était considéré comme un document structuré...
  - $\rightarrow$ à l'exception notable de JeuxDeMots (Mathieu Lafourcade)
  - $\rightarrow$ et de Wiktionary...

Que faire alors ?

- Aller prendre le collaboratif là où il est...
  - Wiktionary...
- Le lexique doit être un graphe
  - RDF
- Le lexique doit être un ouvert et interopérable
  - Sortir du tout XML (illusion de l'interopérabilité)
  - $\rightarrow$ Lexical Linked Data
- Faire émerger les acceptions
  - A partir des données observées

Troisième partie III

DBnary : Lexical Linked Open Data

4 Données liées ouvertes
- RDF facile...
- Linked Data pour les nuls...

5 DBnary
- Qu'est ce que DBnary ?
- Les traductions sont attachées aux sens des mots !
  - Problem
  - Quelle mesure de similarité ?
  - Parameters
  - Extraction of an Endogeneous Gold Standard
- Experimental Protocol
  - Results

4 Données liées ouvertes
- RDF facile...
- Linked Data pour les nuls...

5 DBnary
- Qu'est ce que DBnary ?
- Les traductions sont attachées aux sens des mots !
  - Problem
  - Quelle mesure de similarité ?
  - Parameters
  - Extraction of an Endogeneous Gold Standard
- Experimental Protocol
  - Results

RDF facile...

- RDF
  - L'ingrédient de base est un triplet (tuple)
  - Le sujet est une "Ressource", avec un nom (une URI ou IRI)
  - La relation est une propriétés, avec un nom (idem)
  - L'objet est soit une ressource, soit une valeur immédiate (possiblement typée)
- OWL / RDFS
  - On peut décrire des classes (types de ressources) et formaliser leurs propriétés
  - On peut décrire les propriétés des relations (transitives, symétriques, etc.)
  - Grâce à ces propriétés, on peut faire des raisonnements formels

RDF facile...

- Intérêt de RDF par rapport à XML
  - Il n'existe plus de notion de "document"
  - Plus de frontières entre les données à l'intérieur et les données à l'extérieur
  - Structuration s'ajoute à la donnée originale, plusieurs structurations simultanées possibles
  - Accès aux données "comme on veut", sans avoir à connaître le "chemin" (ex : itérer sur tous les sens...)

4 Données liées ouvertes
- RDF facile...
- Linked Data pour les nuls...

5 DBnary
- Qu'est ce que DBnary ?
- Les traductions sont attachées aux sens des mots !
  - Problem
  - Quelle mesure de similarité ?
  - Parameters
  - Extraction of an Endogeneous Gold Standard
- Experimental Protocol
  - Results

Linked **Open** Data: The 5 star plan

★ Make your data available on the Web under an open license

★★ Make it available as structured data

(Excel sheet instead of image scan of a table)

★★★ Use a non-proprietary format

(CSV file instead of an Excel sheet)

★★★★ Use Linked Data format

(URIs to identify things, RDF to represent data)

★★★★★ Link your data to other people's data to provide context

More: http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

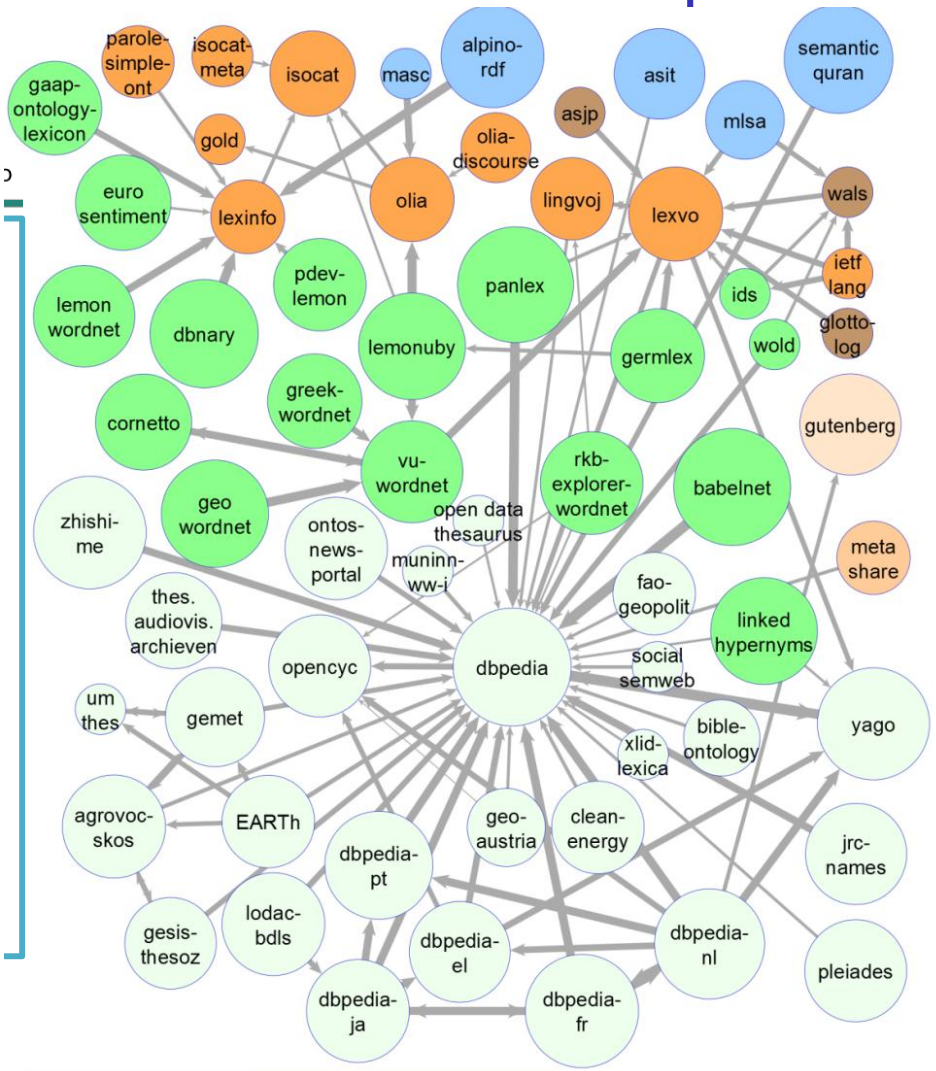
Linked data, et en français, ça donne quoi ?

- Toutes les ressources (les nœuds du graphes) sont déréférençables (URI = URL)
  - ex : `http://kaiko.getalp.org/dbnary/fra/chat`
  - + négociation de contenu
  - . . . si vous êtes un humain avec un navigateur, vous aurez une description lisible en HTML
  - . . . si vous êtes un robot, vous aurez une description lisible en RDF-XML, JSON, N3, TURTLE, ... c'est vous qui décidez...

Linked data, et en français, ça donne quoi ?

- Toutes les ressources (les nœuds du graphes) sont déréférençables (URI = URL)
  - ex : `http://kaiko.getalp.org/dbnary/fra/chat`
  - + négociation de contenu
  - . . . si vous êtes un humain avec un navigateur, vous aurez une description lisible en HTML
  - . . . si vous êtes un robot, vous aurez une description lisible en RDF-XML, JSON, N3, TURTLE, ... c'est vous qui décidez...
- Toutes les ressources sont accessible avec un langage de requêtes SPARQL
  - permet des sélections structurées
  - permet d'exporter les données comme on veut les voir (qui peut le plus, peut le moins...)
- SPARQL permet même de sélectionner des données de différents "datasets"
  - même s'il ne sont pas sur la même machine...
  - vous pouvez travailler avec des données plus grande que votre RAM (et même que votre espace disque) !

The Lexical Linked Open Data Cloud



Linguistic Linked Open Data (LLOD) cloud diagram

May 2014
CC-BY Open Linguistics Working Group
(http://linguistics.okfn.org/llod)

Compiled for the 3rd Workshop on Linked Data in Linguistics (LDL-2014)

4 Données liées ouvertes
- RDF facile...
- Linked Data pour les nuls...

5 DBnary
- Qu'est ce que DBnary ?
- Les traductions sont attachées aux sens des mots !
  - Problem
  - Quelle mesure de similarité ?
  - Parameters
  - Extraction of an Endogeneous Gold Standard
- Experimental Protocol
  - Results

Dbnary : Wiktionary comme graphe lexical

English [edit]

Pronunciation [edit]

- (*UK*) IPA: /kæt/, [kʰæt]
- (*US*) IPA: /kæt/, [kʰæʔ(t̚)], [kʰe̞ə̯t]
- Audio (UK) ▶ 0:00 MENU
- Audio (US) ▶ 0:00 MENU
- Audio (US-Inland North) ▶ 0:00 MENU
- Rhymes: -æt

Etymology 1 [edit]

From Middle English *cat*, *catte*, from Old English *catt* ("male cat") and *catte* ("female cat"), from Late Latin *cattus* ("domestic cat"), from Latin *catta* (c.75 B.C., Martial)[1], from Afro-Asiatic (compare Nubian *kadís*, Berber *kaddîska* 'wildcat'), from Late Egyptian *čaute*,[2] feminine of *čaus* 'jungle cat, African wildcat', from earlier Egyptian *tešau* 'female cat'. Cognate with Scots *cat* ("cat"), Welsh *cath* ("cat"), West Frisian *kat* ("cat"), North Frisian *kåt* ("cat"), Dutch *kat* ("cat"), Low German *Katt, Katte* ("cat"), German *Katze* ("cat"), Danish *kat* ("cat"), Swedish *katt* ("cat"), Icelandic *köttur* ("cat"), Armenian կատու (*katu*, "cat").

Noun [edit]

cat (*plural* **cats**)

1. A domesticated subspecies, *Felis silvestris catus*, of feline animal, commonly kept as a house pet. [from 8th c.]
2. Any similar animal of the family Felidae, which includes lions, tigers, bobcats, etc.
3. A catfish. [quotations ▼]
4. (*offensive*) A spiteful or angry woman. [from earlier 13th c.]
5. An enthusiast or player of jazz.
6. (*slang*) A person (usually male).
7. (*nautical*) A strong tackle used to hoist an anchor to the cathead of a ship.
8. (*chiefly nautical*) *Short form of* cat-o'-nine-tails. [quotations ▼]
9. (*slang*) Any of a variety of earth-moving machines. (from their manufacturer Caterpillar Inc.)
10. (*archaic*) A sturdy merchant sailing vessel (*now only in "catboat"*).
11. (*archaic*, *uncountable*) The game of "trap and ball" (also called "cat and dog").
12. (*archaic*, *uncountable*) The trap of the game of "trap and ball".
13. (*slang*) Prostitute. [from at least early 15th c.]
14. (*slang*, *vulgar*, *African American Vernacular*) A vagina; female external genitalia [quotations ▼]
15. A double tripod (for holding a plate, etc.) with six feet, of which three rest on the ground, in whatever position it is placed.

Wikipedia has an article on:
Cat

A domestic cat (1)

Dbnary : Wiktionary comme graphe lexical

Synonyms [edit]

- (*any member of the* *suborder* *(sometimes* *superfamily)* *Feliformia* *or* Feloidea): feliform ("cat-like" carnivoran), feloid (compare Caniformia, Canoidea)
- (*any member of the* *family* *Felidae*): felid
- (*any member of the* *subfamily* *Felinae*, *genera* Puma, Acinonyx, Lynx, Leopardus, *and* Felis*)*): feline cat, a feline
- (*any member of the subfamily* *Pantherinae*, *genera* Panthera, Uncia *and* Neofelis): pantherine cat, a pantherine
- (*technically, all members of the genus* Panthera): panther (i.e. tiger, lion, jaguar, leopard), (*narrow sense*) panther (i.e. black panther)
- (*any member of the* *extinct* *subfamily* Machairodontinae, *genera* Smilodon, Homotherium, Miomachairodus, *etc.*): Smilodontini, Machairodontini (Homotherini), Metailurini, "saber-toothed cat" (saber-tooth)
- (*domestic species*): housecat, puss, pussy, malkin, kitten, kitty, pussy-cat, mouser, tomcat, grimalkin
- (*man*): bloke (*UK*), chap (*British*), cove (*UK*), dude, fellow, fella, guy
- (*spiteful woman*): bitch
- See also Wikisaurus:cat
- See also Wikisaurus:man

Derived terms [edit]

Terms derived from *cat* in the above senses [show ▼]

Dbnary : Wiktionary comme graphe lexical

Translations [edit]

domestic species	[show ▼]
member of the suborder (or superfamily) Feliformia (Feloidea), "cat-like" carnivorans	[show ▼]
member of the family Felidae	[show ▼]
member of the subfamily Felinae	[show ▼]
member of the subfamily Pantherinae	[show ▼]
member of the extinct subfamily Machairodontinae	[hide ▲]

Select targeted languages

- Esperanto: maĥairodeno (eo)
- French: machairodontiné (fr) *m*, machairodontinés (fr) *pl*
- German: Säbelzahnkatze (de) *f*, Machairodontine (de) *m*, Machairodontine (de) *f*, Machairodontinen (de) *pl*, Machairodontinae (de) *pl*
- Add translation : Preview translation More

Script template: (e.g. Cyrl for Cyrillic, Latn for Latin)

type of fish — *see* **catfish**	
spiteful woman — *see* **bitch**	
jazz enthusiast	[show ▼]
guy, fellow	[show ▼]
strong tackle used to hoist an anchor to the cathead of a ship	[show ▼]
cat-o'-nine-tails — *see* **cat-o'-nine-tails**	
type of boat — *see* **catboat**	
game of "trap and ball" (or "cat and dog")	[show ▼]
the trap in the game of "trap and ball"	[show ▼]

Dbnary : Wiktionary comme graphe lexical

```
==English==
[[Category:English three-letter words]]{{rfc-auto}}
{{wikipedia}}
[[Image:Cat03.jpg|thumb|A domestic cat (1)]]

===Pronunciation===
* {{a|UK}} {{IPA|/kæt/|[kʰæt]}}
* {{a|US}} {{IPA|/kæt/|[kʰæʔ(t̚)]|[kʰeə̯t]}}
* {{audio|En-uk-a cat.ogg|Audio (UK)}}
* {{audio|En-us-cat.ogg|Audio (US)}}
* {{audio|En-us-inlandnorth-cat.ogg|Audio (US-Inland North)}}
* {{rhymes|æt}}

===Etymology 1===
From {{etyl|enm|en}} {{term|cat|lang=enm}}, {{term|catte|lang=enm}}, from {{etyl|ang|en}} {{term|catt||male cat|lang=ang}} and

====Noun====
{{en-noun}}

# A domesticated [[subspecies]], ''[[Felis silvestris catus]]'', of [[feline]] animal, commonly kept as a house [[pet]]. {{defdate|fro
# Any similar animal of the family [[Felidae]], which includes [[lion]]s, [[tiger]]s, bobcats, etc.
# A [[catfish]].
#* '''1913''', [[w:Willa Cather|Willa Cather]], ''[[s:O Pioneers!|O Pioneers!]]'', [[s:O Pioneers!/The Wild Land, II|chapter 2]]:
#*: She missed the fish diet of her own country, and twice every summer she sent the boys to the river, twenty miles to the sou
# {{context|offensive|lang=en}} A spiteful or angry [[woman]]. {{defdate|from earlier 13th c.}}
# An enthusiast or player of [[jazz]].
```

Dbnary : Wiktionary comme graphe lexical

=====Synonyms=====
* {{sense|any member of the [[suborder]] (sometimes [[superfamily]]) [[Feliformia]] or {{taxlink|Feloidea|suborder}}}} [[feliform]
* {{sense|any member of the [[family]] [[Felidae]]}} [[felid]]
* {{sense|any member of the [[subfamily]] [[Felinae]], genera ''[[Puma]]'', ''[[Acinonyx]]'', ''[[Lynx]]'', ''[[Leopardus]]'', and ''[[Fe
* {{sense|any member of the subfamily [[Pantherinae]], genera ''[[Panthera]], [[Uncia]]'' and ''[[Neofelis]]''}} [[pantherine cat]],
* {{sense|technically, all members of the genus ''Panthera''}} [[panther]] (i.e. [[tiger]], [[lion]], [[jaguar]], [[leopard]]), {{qualifier|
* {{sense|any member of the [[extinct]] subfamily ''{{taxlink|Machairodontinae|subfamily}}'', genera {{taxlink|Smilodon|genus|r
* {{sense|domestic species}} [[housecat]], [[puss]], [[pussy]], [[malkin]], [[kitten]], [[kitty]], [[pussy-cat]], [[mouser]], [[tomcat]],
* {{sense|man}} [[bloke]] {{qualifier|UK}}, [[chap]] {{qualifier|British}}, [[cove]] {{qualifier|UK}}, [[dude]], [[fellow]], [[fella]], [[guy]
* {{sense|spiteful woman}} [[bitch]]
* See also [[Wikisaurus:cat]]
* See also [[Wikisaurus:man]]

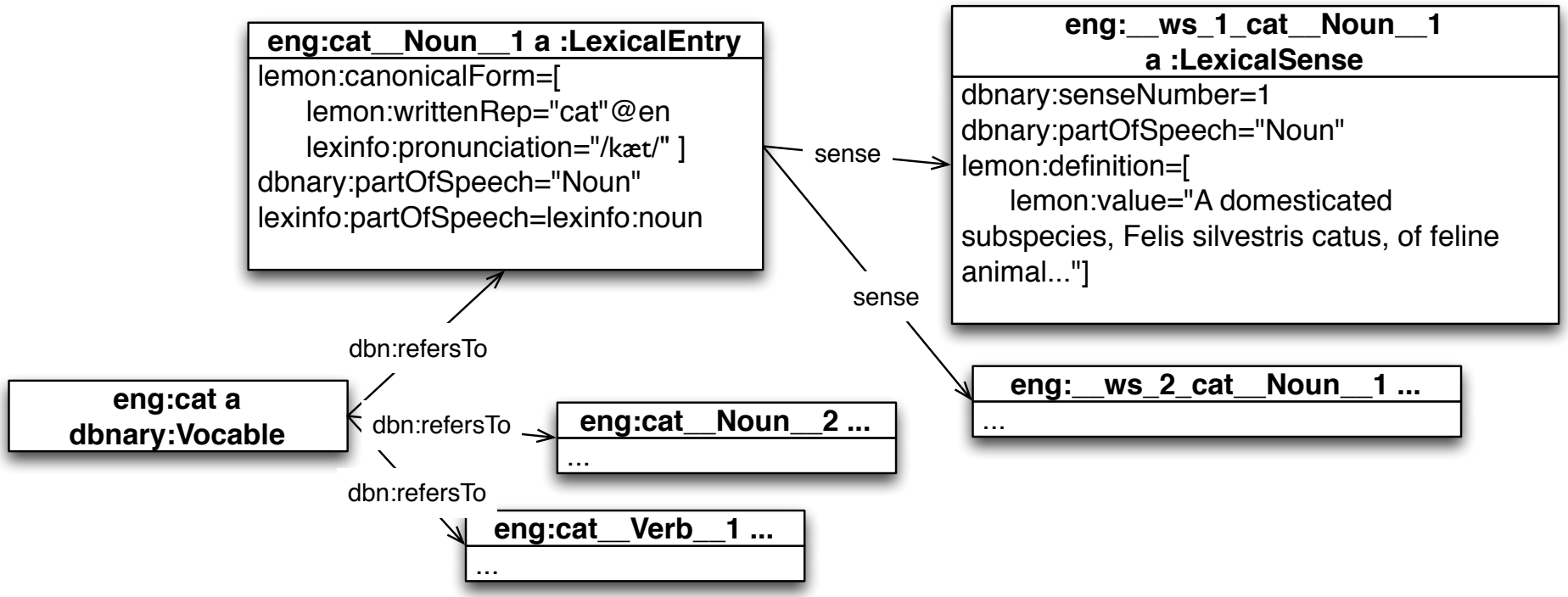
Dbnary : Wiktionary comme graphe lexical

=====Translations=====

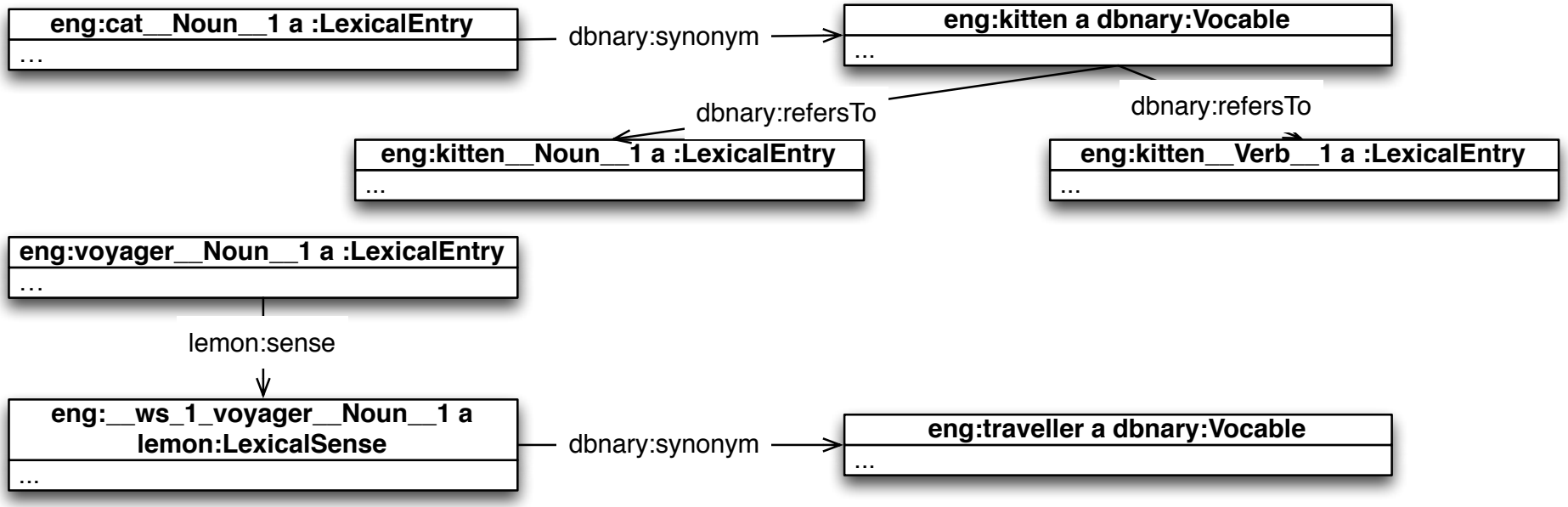
{{trans-top|domestic species}}

* {{trreq|ab}}
* Acehnese: {{tø|ace|mië}}
* Adyghe: {{tø|ady|кIэтыу|sc=Cyrl}}
* Afrikaans: {{t+|af|kat}}
* Ainu: {{tø|ain|チャペ|tr=cape}}
* Akan: [[agyinamoa]] {{n}}
* Albanian: {{t+|sq|mace|f}}
* Alemannic German: {{tø|gsw|Chàtz}}
* Amharic: {{t-|am|ድመት|tr=dəmät|sc=Ethi}}
* Apache:
*: Western Apache: {{tø|apw|gídí}}
* Arabic: {{t+|ar|قط|m|tr=qiṭṭ|sc=Arab}}, {{t+|ar|قطة|f|tr=qíṭṭa|sc=Arab}}
*: Egyptian: {{tø|arz|قط|m|tr='uṭṭ}}, {{tø|arz|قطة|f|tr='uṭṭɑ}}
*: Libyan: {{t-|ar|قطوس|m|tr=gɑṭṭūs|sc=Arab}}, {{t-|ar|قطوسة|f|tr=gɑṭṭūsa|sc=Arab}}
*: Moroccan Arabic: {{tø|ary|مش|tr=mešš}}, {{tø|ary|مشة|f|tr=mešša}}
* Aramaic:
*: Syriac: [[ܩܛܘܠܐ]] (šūnārā') {{m}}, [[ܩܛܘܠܬܐ]] (šūnārtā') {{f}}

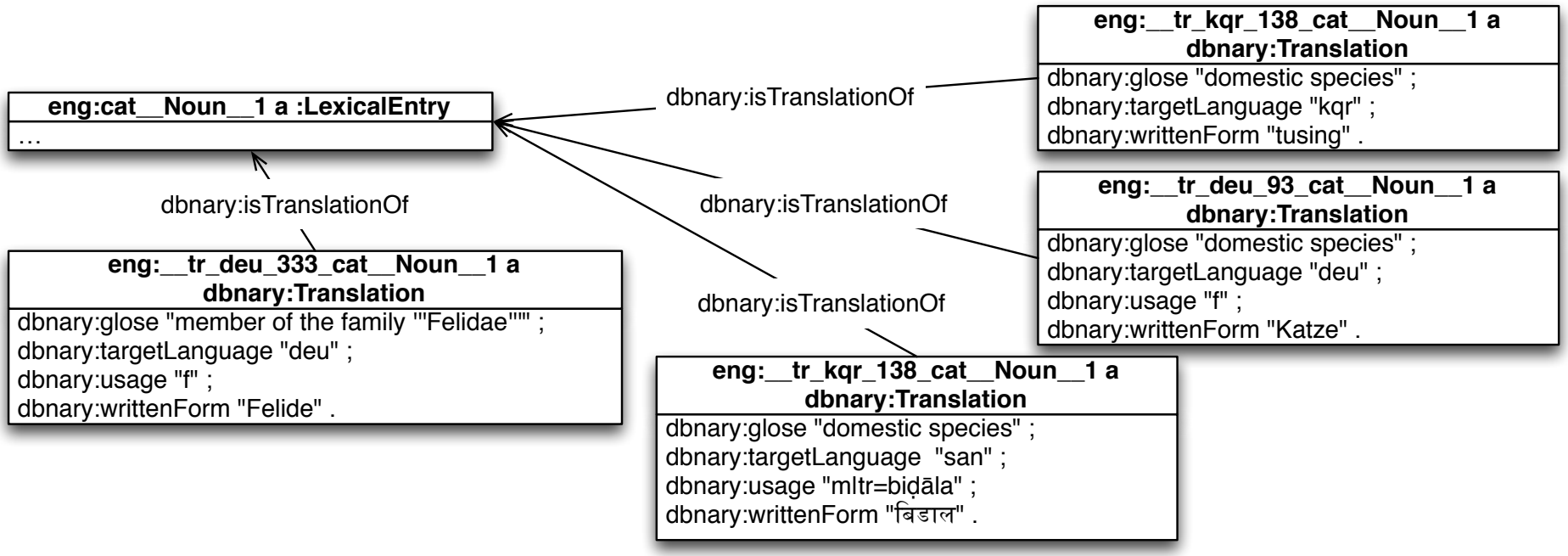
Dbnary : Wiktionary comme graphe lexical



Dbnary : Wiktionary comme graphe lexical



Dbnary : Wiktionary comme graphe lexical



4 Données liées ouvertes
- RDF facile...
- Linked Data pour les nuls...

5 DBnary
- Qu'est ce que DBnary ?
- Les traductions sont attachées aux sens des mots !
  - Problem
  - Quelle mesure de similarité ?
  - Parameters
  - Extraction of an Endogeneous Gold Standard
- Experimental Protocol
  - Results

DBnary ? Késako ?

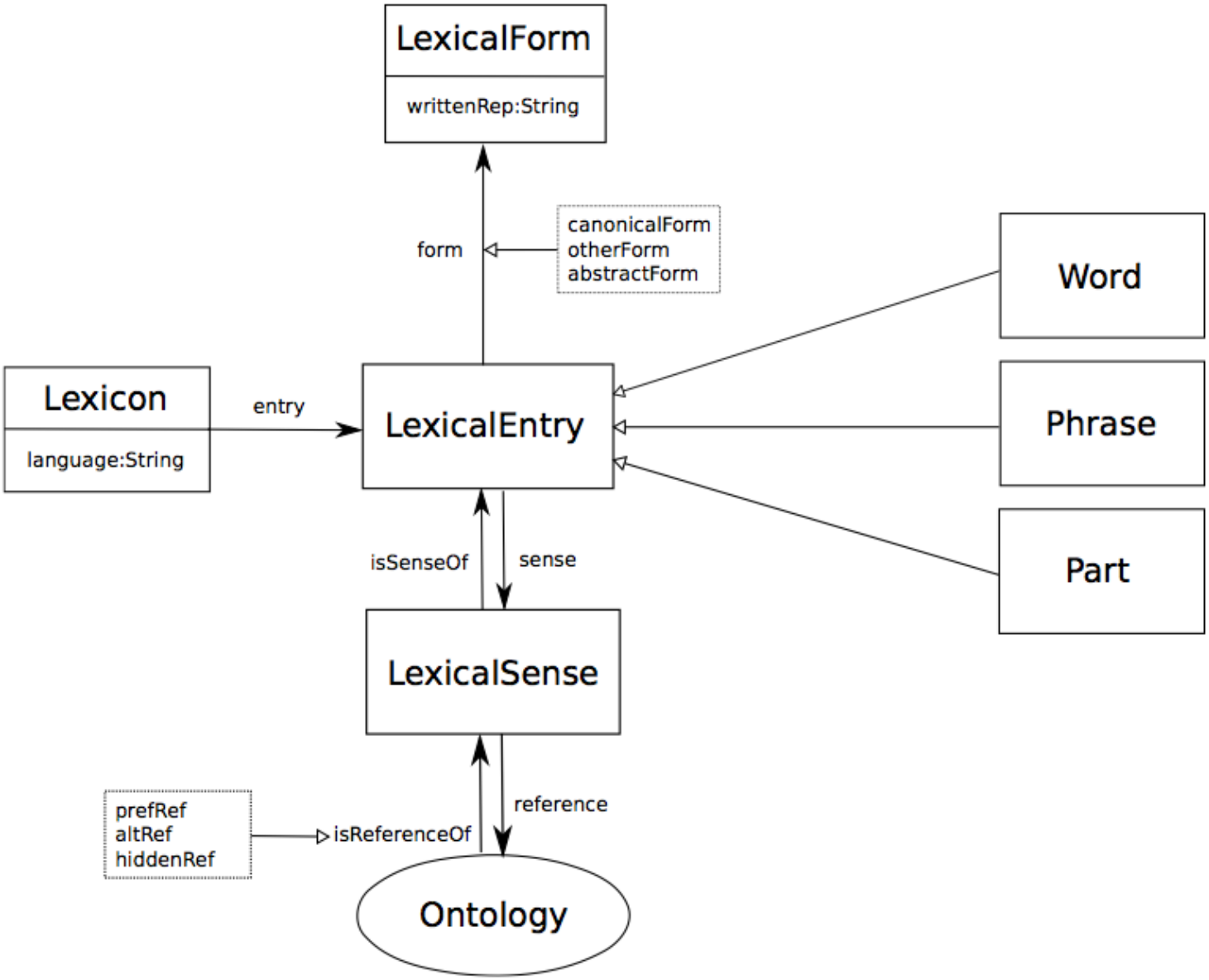
Des données lexicales extraites de 13 éditions différentes des wiktionnaires

DBnary ? Késako ?

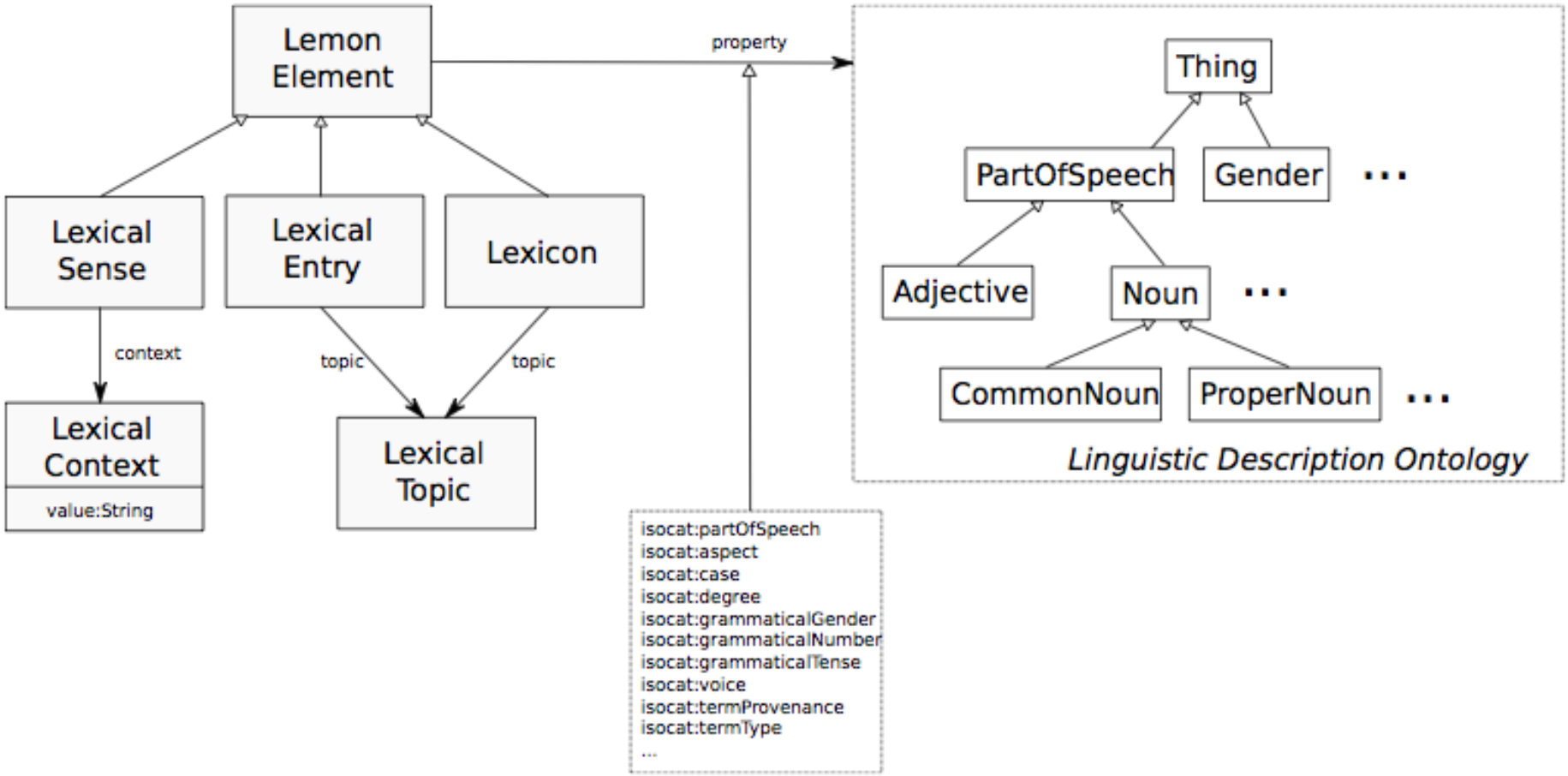
Les données extraites sont disponibles en «Linked Data» (Données liées ouvertes)

→ structurées avec LEMON, LEXINFO et quelques extensions propres à BDnary.

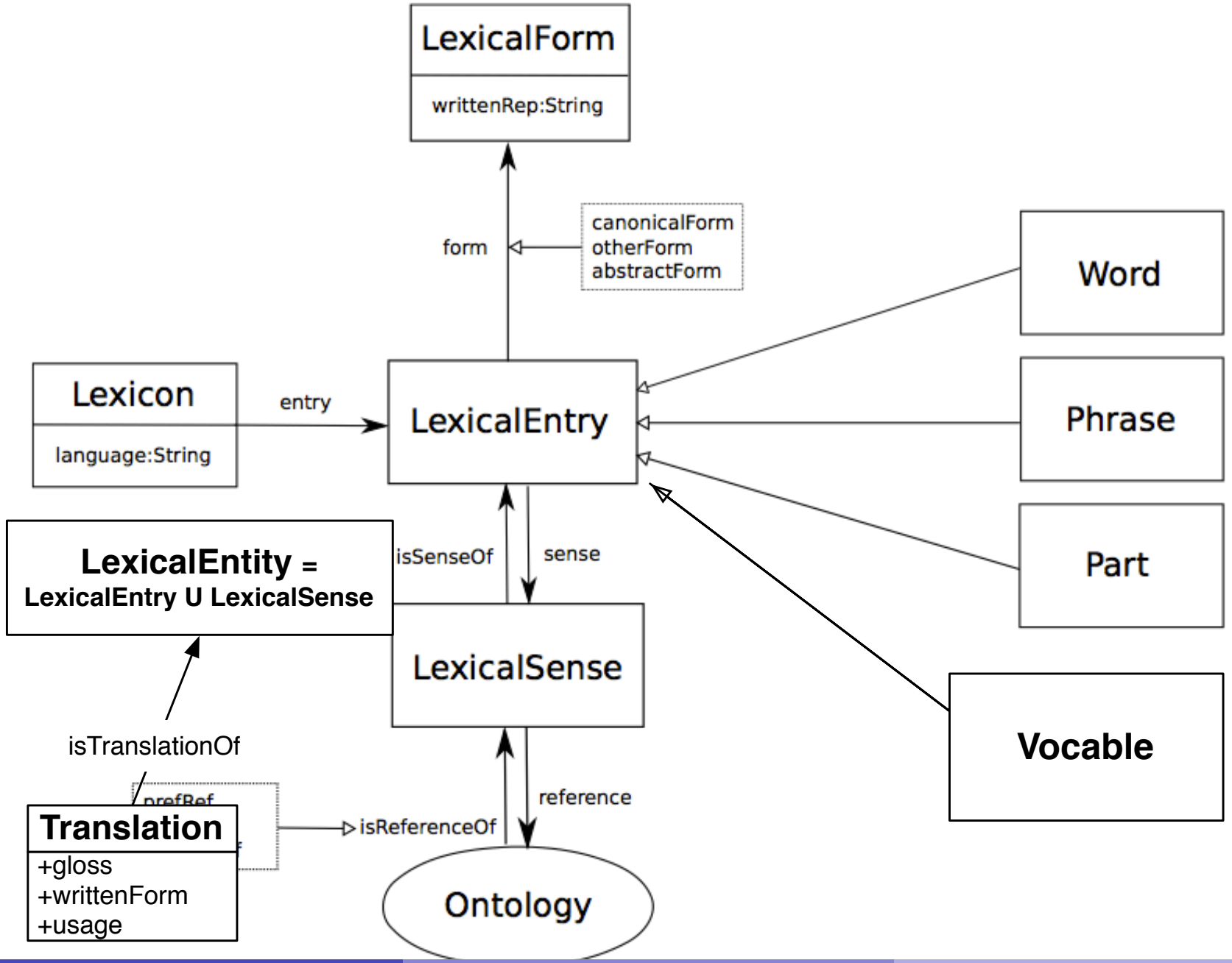
LEMON


LexicalForm
writtenRep:String
form
canonicalForm
otherForm
abstractForm
Word
Lexicon
language:String
entry
LexicalEntry
Phrase
Part
isSenseOf
sense
LexicalSense
prefRef
altRef
hiddenRef
isReferenceOf
reference
Ontology

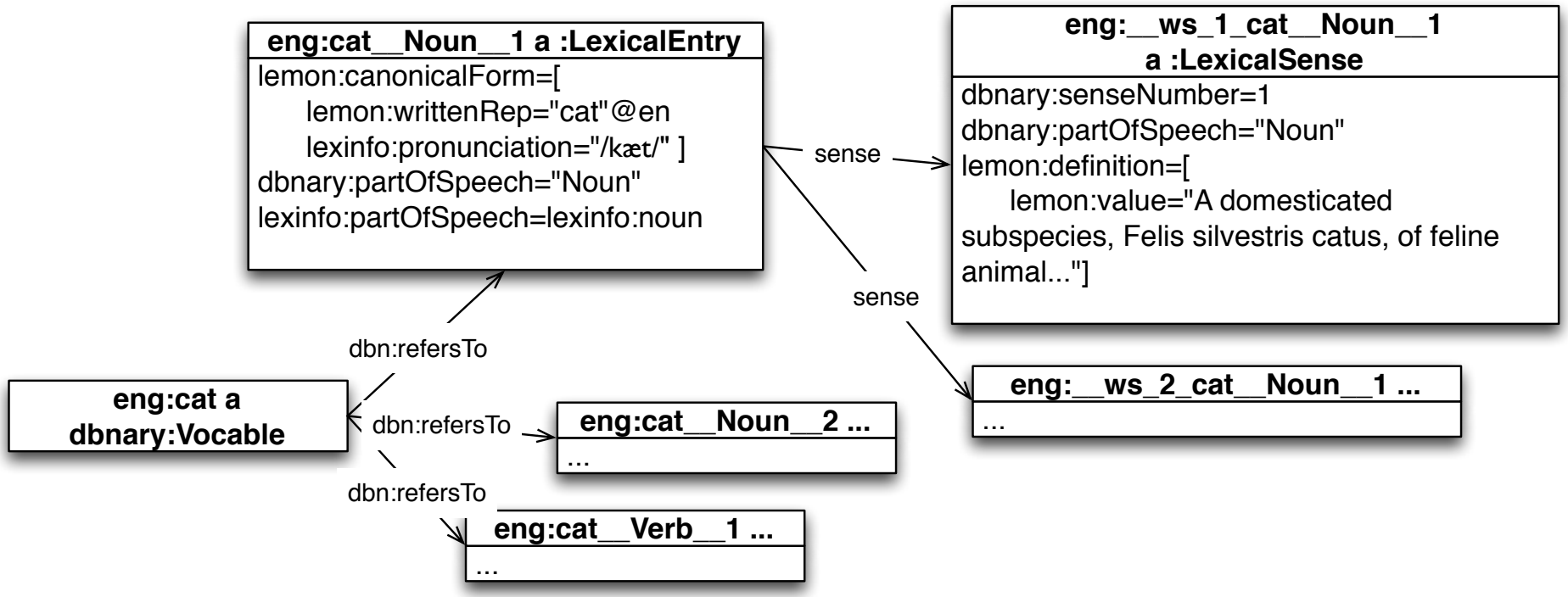
LEMON


Lemon Element
property
Lexical Sense
Lexical Entry
Lexicon
context
topic
topic
Lexical Context
value:String
Lexical Topic
Thing
PartOfSpeech
Gender
Adjective
Noun
CommonNoun
ProperNoun
Linguistic Description Ontology
isocat:partOfSpeech
isocat:aspect
isocat:case
isocat:degree
isocat:grammaticalGender
isocat:grammaticalNumber
isocat:grammaticalTense
isocat:voice
isocat:termProvenance
isocat:termType
...

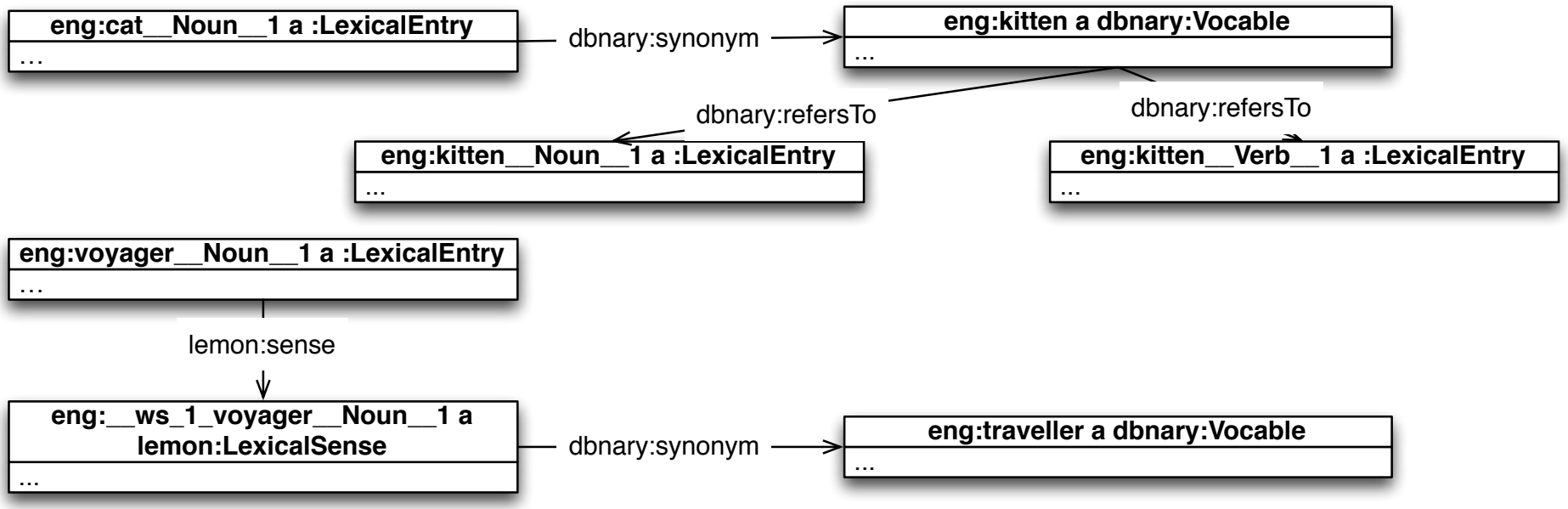
LEMON


LexicalForm
writtenRep:String
form
canonicalForm
otherForm
abstractForm
Word
Lexicon
language:String
entry
LexicalEntry
Phrase
Part
LexicalEntity =
LexicalEntry U LexicalSense
isSenseOf
sense
LexicalSense
Vocable
isTranslationOf
Translation
+gloss
+writtenForm
+usage
isReferenceOf
reference
Ontology

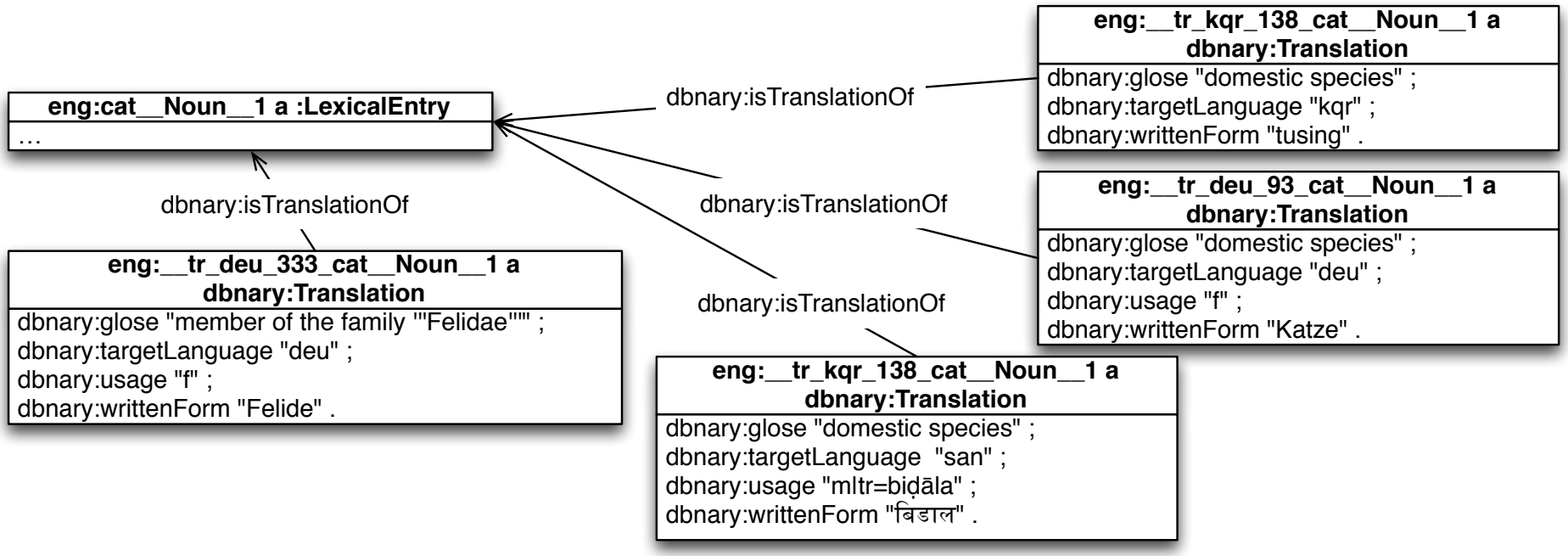
DBnary : exemple



DBnary : exemple



DBnary : exemple



DBnary : exemple

Allons voir les données...

Principe d'extraction

- 13 éditions extraites : en, fr, de, es, it, pt, fi, el, tr, ru, jp, bg, pl
- Extraction à partir des dumps de wikimedia
- Chaque édition met à jour son dump tous les 10-15j
- l'extraction est faite en direct, les fichiers sont accessibles en ligne immédiatement
- le serveur "Linked Data" contient des données plus anciennes

Taille des données (cet été)

	Vocables	Entries	Senses	Translations
English	523,300	562,538	454,285	1,366,385
French	285,497	299,444	388,642	514,525
Greek	242,470	248,107	140,190	58,952
German	212,226	215,808	104,556	396,376
Russian	149,534	152,488	152,394	381,641
Turkish	66,830	67,820	99,148	68,416
Spanish	57,385	60,438	88,022	116,912
Finnish	54,376	55,783	64,931	125,615
Portuguese	42,503	46,553	82,900	269,516
Italian	32,735	35,313	46,367	63,711
Japanese	21,547	26,267	30,720	93,161
Bulgarian	18,792	18,823	18,299	13,925
Total	1,707,195	1,789,382	1,670,454	3,469,135

TABLE: Number of resources by type and language, sorted by number of lexical entries.

Taille des données (cet été)

	syn	ant	hyper	hypo	holo	mero
German	123,550	55,158	95,962	81,467	0	0
English	32,147	7,056	1,210	1,234	0	114
French	32,089	6,929	9,673	3,776	1,914	978
Russian	25,844	10,134	42,955	5,250	0	0
Bulgarian	17,623	34	0	0	0	0
Spanish	15,562	1,602	784	565	0	0
Italian	10,148	3,718	0	0	0	0
Greek	5,439	1,553	0	0	0	0
Japanese	4,002	1,631	10	18	0	0
Portuguese	3,395	621	6	4	0	0
Turkish	3,219	242	484	166	0	0
Finnish	2,654	0	0	0	0	0
Total	275,672	88,678	151,084	92,480	1,914	1,092

TABLE: Number of lexico-semantic relations. Languages are sorted according to the number of synonym.

Taille des données (cet été)

Source/Target	deu	ell	eng	fin	fra	ita	por	ru
eng	62501	23794	1	74938	57959	37467	30256	7483
fra	34608	7063	74687	7589	12	18806	17784	778
deu	0	2675	81015	4947	67143	41485	8872	1735
rus	23056	3295	48559	3966	14776	12643	5567	
ell	2242	2	10090	1056	8436	1470	1149	131
fin	8046	918	30103	0	6700	3856	2196	799
por	7000	2816	11284	4607	8720	7096	4	439
ita	4619	506	17539	925	4461	75	1219	93

TABLE: Number of translations from/to the 8 currently extracted languages. Source languages are sorted according to their number of lexical entries. Target languages are sorted by their ISO 639-3 language code. The number of different target languages is also given.

Taille des données (cet été)

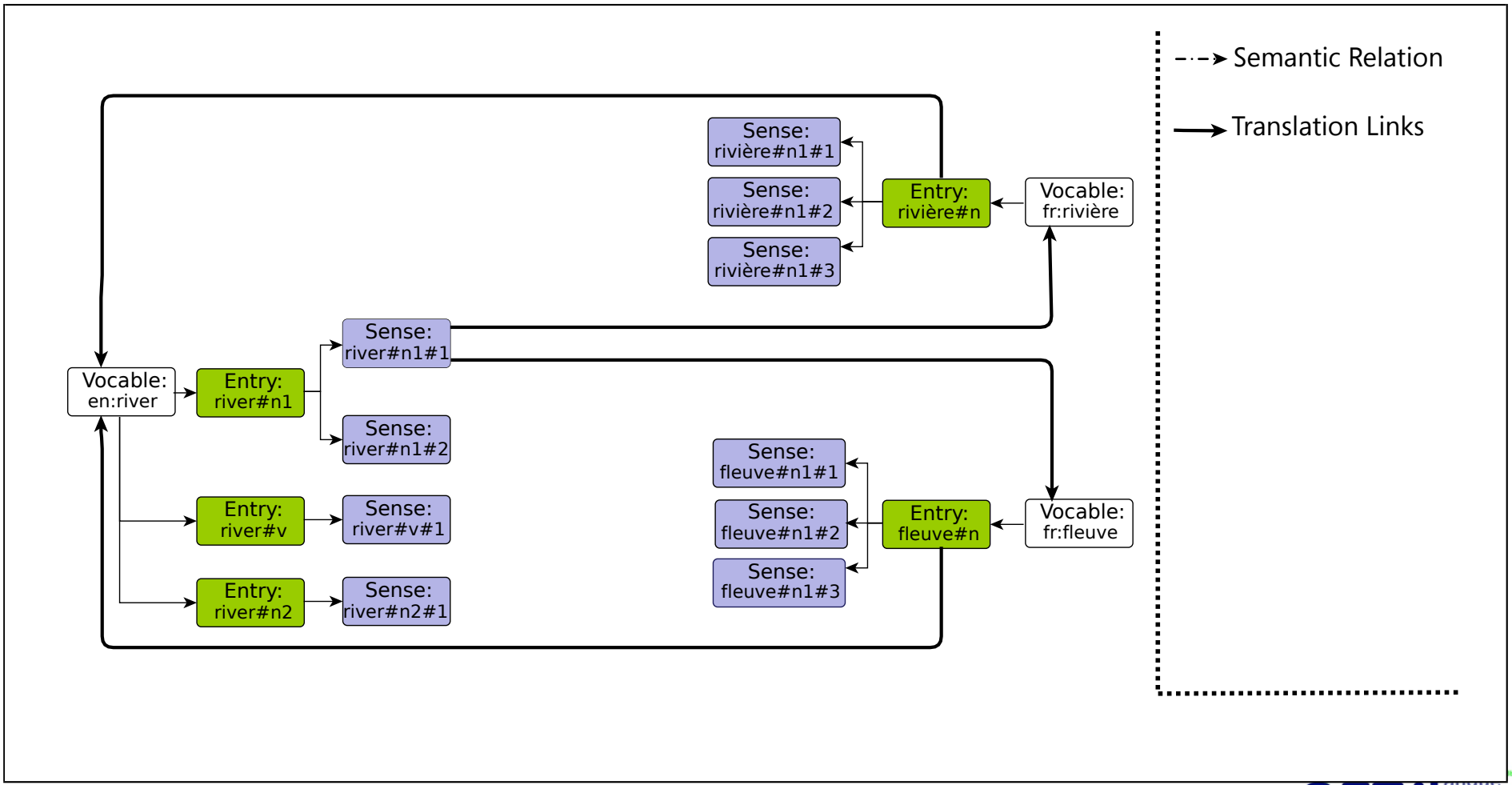
Source/Target	por	rus	others	Total	# of lang
eng	30256	74837	764710	1126463	1143
fra	17784	7783	296624	464956	952
deu	8872	17354	248401	471892	355
rus	5567	0	206709	318571	490
ell	1149	1315	29892	55652	246
fin	2196	7997	58912	118728	329
por	4	4396	179142	225065	695
ita	1219	938	27514	57796	315

TABLE: Number of translations from/to the 8 currently extracted languages. Source languages are sorted according to their number of lexical entries. Target languages are sorted by their ISO 639-3 language code. The number of different target languages is also given.

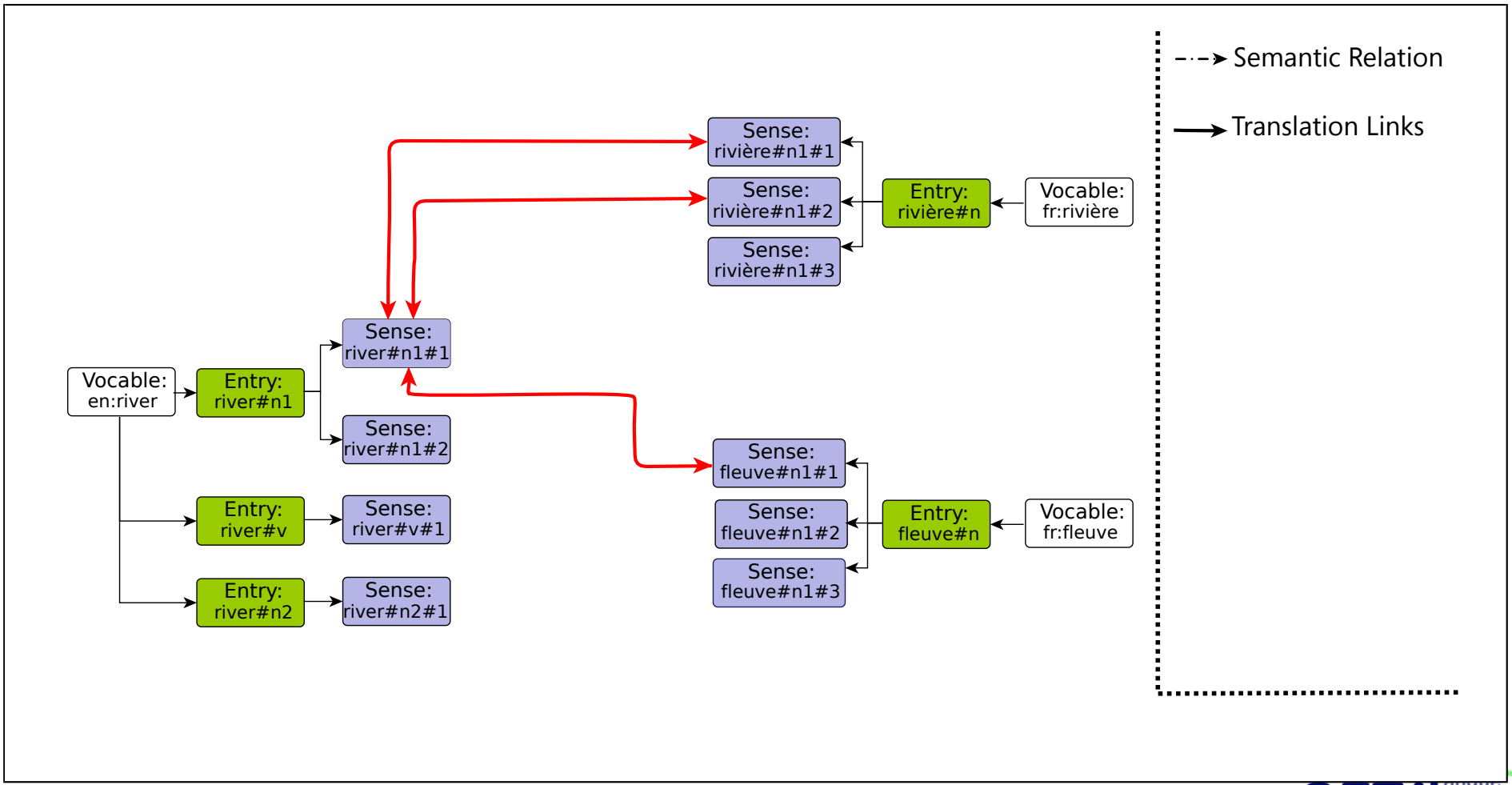
4 Données liées ouvertes
- RDF facile...
- Linked Data pour les nuls...

5 DBnary
- Qu'est ce que DBnary ?
- Les traductions sont attachées aux sens des mots !
  - Problem
  - Quelle mesure de similarité ?
  - Parameters
  - Extraction of an Endogeneous Gold Standard
- Experimental Protocol
  - Results

The problem


Semantic Relation
Translation Links
Sense: rivière#n1#1
Sense: rivière#n1#2
Sense: rivière#n1#3
Entry: rivière#n
Vocable: fr:rivière
Sense: river#n1#1
Vocable: en:river
Entry: river#n1
Sense: river#n1#2
Sense: fleuve#n1#1
Sense: fleuve#n1#2
Sense: fleuve#n1#3
Entry: fleuve#n
Vocable: fr:fleuve
Entry: river#v
Sense: river#v#1
Entry: river#n2
Sense: river#n2#1

The problem


Semantic Relation
Translation Links
Sense: rivière#n1#1
Sense: rivière#n1#2
Sense: rivière#n1#3
Entry: rivière#n
Vocable: fr:rivière
Sense: river#n1#1
Vocable: en:river
Entry: river#n1
Sense: river#n1#2
Sense: fleuve#n1#1
Sense: fleuve#n1#2
Sense: fleuve#n1#3
Entry: fleuve#n
Vocable: fr:fleuve
Entry: river#v
Sense: river#v#1
Entry: river#n2
Sense: river#n2#1

Some Cues

As an example, the English `LexicalEntry` frog contains 8 word senses, defined as follows :

1. A small tailless amphibian of the order Anura that typically hops
2. The part of a violin bow (or that of other similar string instruments such as the viola, cello and contrabass) located at the end held by the player, to which the horsehair is attached
3. (Cockney rhyming slang) Road. Shorter, more common form of frog and toad
4. The depression in the upper face of a pressed or handmade clay brick
5. An organ on the bottom of a horse's hoof that assists in the circulation of blood
6. The part of a railway switch or turnout where the running-rails cross (from the resemblance to the frog in a horse's hoof)
7. An oblong cloak button, covered with netted thread, and fastening into a loop instead of a button hole.
8. The loop of the scabbard of a bayonet or sword.

Translations of this entry are divided in 4 groups corresponding to : "amphibian", "end of a string instrument's bow", "organ in a horse's foot" and "part of a railway". This is what we call the **glosses**.

Different languages, different cues

- English language edition uses glosses to annotate translations.
- German language edition uses the sense number to annotate translations.
- French language edition uses sometimes a gloss, sometime a sense number.

Glosses vs Sense Numbers

Language	Translations	w. gloss	w. sense num.	w. both
English	1317545	1288667 (97.81%)	515	515
Finnish	121278	120329 (99.22%)	115949	115550
French	504061	135612 (26.90%)	28821	28114
German	388630	3101 (0.80%)	385452	0
Greek	56638	8368 (14.77%)	12	12
Italian	62546	0 (0.00%)	0	0
Japanese	85606	20686 (24.16%)	4148	2512
Portuguese	267048	72339 (27.09%)	71734	69172
Russian	360016	150985 (41.94%)	115	0
Turkish	66290	585 (0.88%)	52901	138

Where do we go from here ?

- For some languages, we only need to match sense numbers (e.g. Turkish, German, ...)
- for other languages, we cannot do anything for now (e.g. Italian)
- for others, we need to match a gloss to the definition it rephrases (e.g. English, French, ...)
- and some editions are kind enough to offer us a Gold Standard on this very specific task...

Selecting a Similarity Measure

Cues available

- Some translation sources contain glosses that reprise a part of the definition of the corresponding senses
- We can compare the gloss to sense definitions

Constraints

- Gloss overlap measure
- No lemmatisation/stemming $\Rightarrow$ approximate matching

Selecting a Similarity Measure

Possibilities

- Text similarity → Strings, no word distinctions
- Simple overlap → Few matches on surface forms
- Hybrid measures ⇒ Combine word overlap + Text similarity

Hybrid measure

- An overlap measure (Lesk, Jaccard, Tverski, etc.)
- Set cardinality ⇔ Sum of pairwise word textual similarities
- What is the most suitable textual similarity measure ?

Our Similarity Measure

Tversky index

$$\frac{|d_1 \cap d_2|}{|d_1 \cap d_2| + \alpha|d_1 - d_2| + \beta|d_2 - d_1|}$$

Hybrid Tverski

$$|d_1 \cap d_2| \longrightarrow \sum_{(x,y) \in d_1 \times d_2} sim(x, y)$$

Parameters to Estimate

- δ — Similarity value range around the best disambiguation
- α & β — Relative importance of the differences in one or the other set of words. If we want a Tverski index that remains between 0 and 1, then we must have $\alpha = 1 - \beta$.
- *sim* — The similarity measure to use, we have the choice between
  - Jaro-Winkler (FTiJW)
  - Scaled-Levenshtein distance (FTiLs)
  - Monge-Elkan (FTiME)
  - Normalized Longest Common Substring (FTiLcss)
  - None (Ti)

Extraction of an Endogeneous Gold Standard

- Translation links with glosses ∩ Translation links with sense numbers
  - Gold standard – entries built from sense numbers
  - Algorithm – sense numbers removed, only glosses considered

- Three language editions with sufficient data :
  - Finnish – 115, 550
  - Portuguese – 69, 172
  - French – 28, 114

- Gold standard and algorithm output in Trec_eval format

4 Données liées ouvertes
- RDF facile...
- Linked Data pour les nuls...

5 DBnary
- Qu'est ce que DBnary ?
- Les traductions sont attachées aux sens des mots !
  - Problem
  - Quelle mesure de similarité ?
  - Parameters
  - Extraction of an Endogeneous Gold Standard
- Experimental Protocol
  - Results

Experimental Protocol

1. Estimate independent parameters — α, β, *sim*
2. Evaluate the results w.r.t. dependent parameter — δ
3. Compare results to textual similarity hybrid measures w.r.t δ
   - Level 2 Jaro-Winkler (L2JW)
   - Level 2 Scaled Levenstein distance (L2Ls)
   - Level 2 Monge-Elkan (L2ME)

Results w.r.t δ

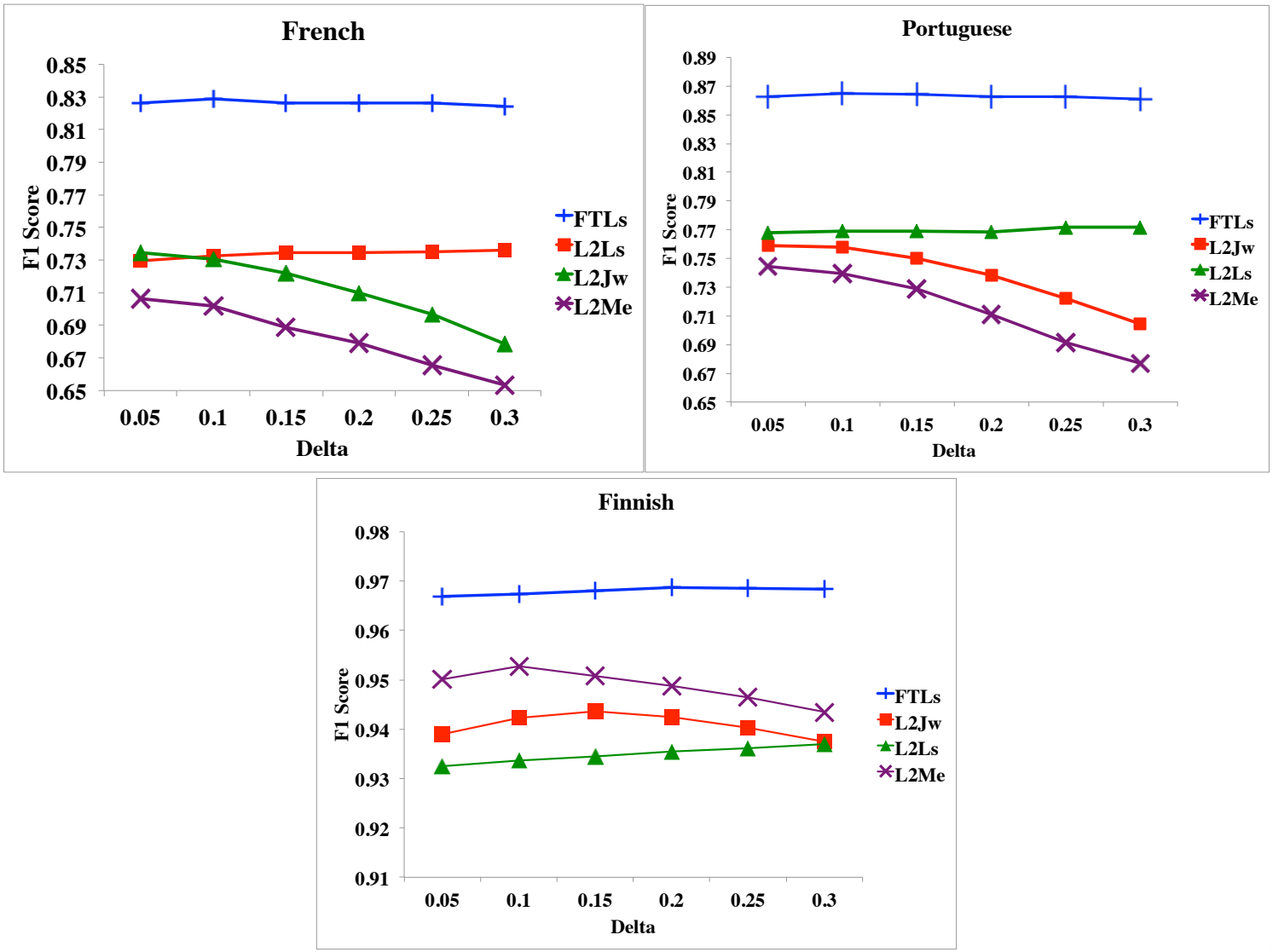


FIGURE: F1 score against delta for our measure and other Level 2 Measures.

Results w.r.t δ

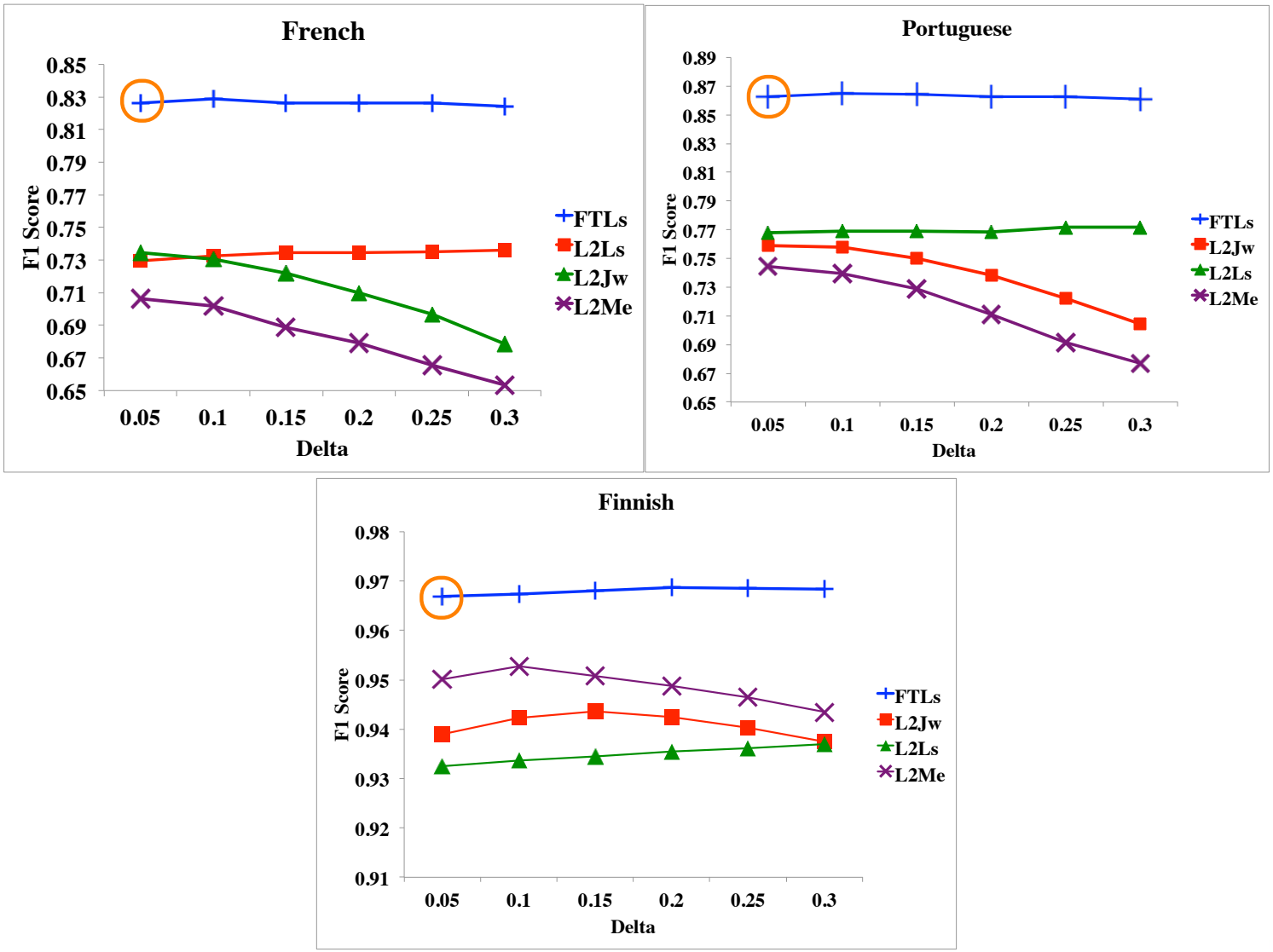


FIGURE: F1 score against delta for our measure and other Level 2 Measures.

Final Results — Relations added to the dataset

	added	P	R	F1	MFS F1	Random
Bulgarian	10,068					
German	388,988					
Greek	8,506					
English	1,299,299					
Spanish	61,079					
Finnish	121,660	0.9642	0.9777	0.9687	0.7218	0.7962
French	136,685	0.8267	0.8313	0.8263	0.3542	0.3767
Italian	0					
Japanese	22,229					
Portuguese	74,426	0.8572	0.8814	0.8651	0.2397	0.3103
Russian	153,485					
Turkish	51,791					

Quatrième partie IV

6 Et maintenant ?

Et maintenant ?

- Lier les données de DBnary à des descriptions linguistiques génériques
  - Morphologie (en cours pour l'allemand et le français)
  - Syntaxe ?
  - fonctions lexicales (Alain, es-tu là ?)
  - ...

Et maintenant ?

- Lier les données de DBnary à des descriptions linguistiques génériques
  - Morphologie (en cours pour l'allemand et le français)
  - Syntaxe ?
  - fonctions lexicales (Alain, es-tu là ?)
  - ...
- Faire émerger des acceptions interlingues à partir des traductions
  - Une acception interlingue est une classe d'équivalence
  - la relation d'équivalence relie des sens de différentes langues
  - elle est inconnue, mais elle est observable au travers des relations de traduction...
  - $\rightarrow$ thèse d'Andon Tchechmedjiev

Et maintenant ?

- Lier les données de DBnary à des descriptions linguistiques génériques
  - Morphologie (en cours pour l'allemand et le français)
  - Syntaxe ?
  - fonctions lexicales (Alain, es-tu là ?)
  - ...
- Faire émerger des acceptions interlingues à partir des traductions
  - Une acception interlingue est une classe d'équivalence
  - la relation d'équivalence relie des sens de différentes langues
  - elle est inconnue, mais elle est observable au travers des relations de traduction...
  - $\rightarrow$ thèse d'Andon Tchechmedjiev
- Comment aborder la "longue traîne" (la myriade de langues peu/moyennement dotées) ?