Methodological provisions in the construction of idiom resources

Presentation · November 2006

1 author:

Eric Guy Claude Laporte
University of Paris-Est
92 PUBLICATIONS **197** CITATIONS

SEE PROFILE

Collocations and idioms 2006:
Linguistic, computational and psycholinguistic perspectives
Berlin, Nov. 3, 2006

Methodological provisions in the construction of idiom resources

Eric Laporte
Institut Gaspard-Monge
Université de Marne-la-Vallée
France

http://www-igm.univ-mlv.fr/~laporte/

Why construct resources describing idioms?

Defining objectives of quality
- accuracy
- coverage

Data- or computer-based provisions
- corpus attestations
- statistical analyses
- golden-standard-based evaluation

Human-based provisions
- objective
- introspective

Why construct resources describing idioms?

Linguistic interest
Idioms make up a large part of languages

Computer applications
Text analysis for information retrieval, information extraction, translation...
Text generation

What kinds of idioms?

Verbal	*make ends meet*
Adverbial	*in the long term*
Nominal	*American coffee*
Adjectival	*rough and ready*
Prepositional ph.	*in a hurry*

Not support verb constructions *make a decision*

1. Defining objectives of quality

Goals and methodological provisions must be adapted to each other

Provisions depend on goals

Provisions are responses to goal-specific risks

Example

Objective: know idioms in 1st century AD Latin

Provision: gather 1st century AD Latin text

More ambition, more methodological provisions

Compatibility between objectives and provisions

Example

Objective: know idioms in 1st century AD Latin

Provision: human control over acceptability of idioms

Trade-off between ambition and provisions

Defining objectives of quality

General objective of quality
Conformity with linguistic reality
Inclusion of all relevant information

Realistic goals
Already attained for some languages

Selected objectives of quality: **accuracy**

Complementarity with grammar

The salmons swim up the river	grammar
John drank up his beer	grammar
Mike gave up the piano	idiom resource

Compositional: grammar
Non-compositional: idiom resources
In fact, idioms also require a grammar

Formalization of description

Conventional dictionaries, second-language grammars...
are interesting but not formalized enough for computer exploitation

Selected objectives of quality: **consistency between intended and actual coverage**

Independence from authors' idiolects

au petit bonheur la chance (Fr.)
au petit bonheur de la chance (my idiolect)
discovered through paper reviewing

Geographical limits

Luc amuse le temps (Québec)
**Luc amuse le temps* (France)

Limits with respect to language plays

Inclusion of variants

Recall or completeness vs. silence or undergeneration
Precision vs. noise or overgeneration

Intended and actual coverage

Completeness vs. undergeneration
Examples of undergeneration

Neglecting variants
in the long term
in the very long term

Consider an idiom as compositional (i.e. taken into account by grammar)
pomme de terre (recent conversation with a linguist)

Intended and actual coverage

Precision vs. overgeneration

Inclusion of obsolete idioms (out-of-date dictionaries)
It rains cats and dogs (?)

Admission of unacceptable variants
John is on the verge of giving up again
**John is on a new verge of giving up*

Checking lemmas, not inflected forms

Il faut voir les choses en face	idiomatic meaning
Il faut voir la chose en face	no idiomatic meaning

Intended and actual coverage

The linguistic notion underlying over- and undergeneration is obviously that of **constraints**

Example
Co-reference of possessives
Leurs hôtes préviennent <u>leurs</u> désirs
not necessarily co-referent to subject
Leurs hôtes reprennent <u>leurs</u> esprits
co-referent to subject
(cf. *lose one's temper*)

Intended and actual coverage

Goals with respect to language plays

Examples: creative reworking of lexicalised metaphors

John spilled the beans lexicalised

John spilled the beans of their relationship lexicalised

John spilled coffee on the bed and the beans of their relationship with it creative

La direction a jeté le bébé avec l'eau du bain lexicalised

La direction a jeté le bébé de la qualité avec l'eau du bain de la formation creative

Intended and actual coverage

A realistic goal

Include: Fully lexicalised forms

Limits of variation of fully lexicalised forms

Exclude: Creative reworking

A basis for future studies about creative reworking

Intended and actual coverage

Syntactic variants

Someone spilled the beans	idiomatic
The beans were spilled	idiomatic
The beans spilled	not idiomatic

A realistic goal

Describe idiomatic variants of idioms
Link all variants of each idiom
Ex. Freckleton 1985, Machonis 1985

A common overgeneralization

A frequent base form, unfrequent variants

Luc n'a pas été gâté par la nature	more frequent
La nature n'a pas gâté Luc	less frequent, active

Other objectives of quality

Less relevant

- psychological plausibility of description
- etymology
- ...

2. Data- and computer-based provisions

Corpus linguistics
A reaction to biased introspective linguistics:

- normativity
- idiolect generalization
- tendency to disregard contexts
- reliance on incomplete conventional dictionaries
- necessity of updates

Convergence with computational linguistics
Automatization of corpus linguistics

Corpus attestations

Attestations give information about existence and frequency of idioms (example: the 'Collocations in the German Language' project)
Balanced corpora
Annotated corpora
The web as corpus (example: the BFQS project)

Recognising the limits of language plays
Context: headlines, advertisement...
Requires intuition also

Corpus attestations

Concordancers

Most corpus linguists use concordancers without lexicons

Unitex, an open-source generator of lemmatized concordances from raw corpora

http://igm.univ-mlv.fr/~unitex

Contains lexicons produced through introspective approaches

Corpus attestations

Results
Conventional dictionaries (e.g. COBUILD) for human users

Problem
No attestations of unacceptability

petite cuillère 'tea spoon' absent from a large Canadian corpus of French texts

Corpus-dependent information about frequency can be in contradiction with real language use (Garrigues 1993)

Statistical analysis

Can be seen as a methodological provision against subjectivity
For many researchers, other motivations: more fun ('Manual construction of resources is tedious'), better salaries?...

Example
Statistical attraction as a sign of frozenness
Similarity of contexts as a sign of semantic proximity
More efficient on technical terms than on verbal idioms

Human revision
Required (methodological provisions: human-based, part 3)

Statistical analysis

Problems

Quality of results of automatic analysis of natural language:

- shallow parsing
- small tagsets
- incomplete data about sense distinctions

Unfrequent idioms are a challenge (e.g. variations, constraints)

Detection of properties: semantic properties, creative reworking of idioms

Statistical analysis

Results

Lists only: properties (variants, constraints) still largely out of reach

Usually not made available
Terminological lists placed on the market

Golden-standard evaluation

Evaluation of an idiom extractor
Manual annotation of a sub-corpus (golden standard)
Comparison with results of automatic extraction

Problems
Golden standards for idioms are small and rare
Little communication about methodological problems in building them (human-based provisions)

Lexicon-Grammar of idioms as Golden standard

A manually constructed Lexicon-Grammar of French idioms
Authors: Maurice Gross, Laurence Danlos
10.000 entries
Made available on line in 2006
http://infolingu.univ-mlv.fr/english

Users
Use as golden standard
Do not be scared by so much information, you can use only the lists if you prefer so

Users and descriptive linguists
Constructive criticism is welcome

Lexicon-Grammar of idioms

N0 =: Nhum	N0 =: N-hum	Ppv	<ENT>	N0 V	<ENT>Det1	N0 V N1 Prep N2	<ENT>N1	N1 =: Npc	[passif]
+	+	<E>	détendre	-	la	-	atmosphère	-	+
+	-	<E>	détenir	-	la	-	vérité	-	+
+	-	<E>	déterrer	-	la	-	hache de la guerre	-	+
+	+	<E>	détourner	-	la	-	conversation	-	+
+	-	<E>	détourner	-	la	-	tête	+	-
+	-	<E>	devancer	-	le	-	appel	-	+
+	-	se	dévisser	-	le	-	cou	+	-
+	+	<E>	dévorer	-	les	-	distances	-	-
+	+	<E>	dévorer	-	les	-	kilomètres	-	-
+	+	<E>	dévorer	-	la	-	route	-	-

3. Human-based provisions

Objective:	psycholinguistic experiments
Introspective:	avoid preconceptions
	native linguists
	mutual control
	time limitation
	readability of resources
	formal criteria
	differential semantic judgment

Psycholinguistic experiments

A reaction to biased introspective linguistics
Separate informant from scientist
Control age, sex, origin, number... of informants

Examples
Recognising idioms as such
Paraphrasing idioms

Psycholinguistic experiments

Drawbacks

Typical time required by an experiment on 20 forms: 2 months

Extrapolated velocity of construction of resources: 40 lexical entries/year (counting 3 forms/entry)

Usually, the idioms need to be known beforehand

Not applicable for comprehensive resources

Human-based provisions: introspective

Specific solutions to the biases of introspective linguistics
Methodology and actual description simultaneously

European tradition
Lexis/grammar interaction
Description of idioms: 1980-now
http://infolingu.univ-mlv.fr/english

American tradition
Wordnet

Avoid preconceptions (1/2)

Preconception 1
'Manual construction of language resources is too difficult'
'Manual construction of resources is error-prone'
Frequently read in (peer-reviewed) computer scientists' papers

The quality of manually constructed resources depends on the background, skills, training and effort of authors
Cf. software

Dysfunctioning of scientific democracy in a case of multi-disciplinarity

At stake: the future of the institutions around the world that train people to construct high-quality language resources

Avoid preconceptions (2/2)

Preconception 2
'Descriptive linguistics is not difficult enough to be interesting'
'Descriptive linguistics does not require much skill'
'Making lists is not the point'

In fact, results of descriptive linguistics are basic information for theoretical linguistics and for computer applications

Native linguists

Native linguists are much better than non-native ones at
- taking into account sense distinctions
- inserting idioms in relevant sentences (this ensures that context is taken into account)
- taking into account semantic properties

Example
La défense a cité un témoin
témoin can have co-referents
Le patron a chié une pendule (not an elegant phrase)
pendule cannot have co-referents

Native linguists

Drawbacks

Results depend on skill, training and effort of the linguist

Not applicable to languages without native speakers with higher education

Not applicable to extinct languages

Mutual control

An idiom resource should be built by a team

Examples
Gross' Lexicon-Grammar of French verbal idioms
Most idioms were listed during the meetings of construction of the Lexicon-Grammar of French verbs (5 linguists)

The Belgium/France/Québec/Switzerland (BFQS) project
Differences between idioms in these 4 varieties of French (4 to 6 linguists)

The BFQS project

Idiom	B	F	Q	S	Paraphrase	Example
Amuser à des riens (s')	+	+	+	+	Se distraire avec des futilités	Il est comme un petit enfant, il s'amuse à des riens.
Amuser à un rien (s')	+	!	-	+	Se distraire avec des futilités	
Amuser bien (s')	+	-	-	-	Se plaire quelque part	Est-ce que tu t'amuses bien dans ton nouvel appartement ?
Amuser la galerie	+	+	+	+	Distraire l'assistance	"Lorsqu'il était petit, il amusait la galerie avec ses mimiques, ses blagues : un acteur était né. " (www)
Amuser le tapis	-	+	-	+	Distraire l'assistance	"Raffarin veut-il amuser le tapis ? Après tout, pourquoi pas, mais la situation dramatique de la France mérite mieux." (www)
Amuser le temps	-	-	+	-	Faire passer le temps	Pierre n'a rien fait de la journée. De plus en plus, j'ai l'impression qu'il amuse le temps.

The BFQS project

1. Make a separate list for each variety
2. Compare lists

Comparison requires meetings
If an idiom is not in the F list, the F author can have missed it
If an idiom in the B list is not understood by the F author, it is considered evidence that it does not belong to the F variety
Intermediate case: passively understood, not actively used
If an idiom in the B list is understood by the F author, compare interpretations, they can be different

Mutual control

Drawback
Cost: several years of weekly or monthly meetings

The grant for the BFQS project will cover only a part of publication costs

Readability of description

Goal: facilitate critical reviewing, update of resources

Example
Table representation
Rows: lexical items
Columns: structure and properties
Open-source software: HOOP (Sastre 2006)

Density of representation
Number of lexical items on the same screen or page
Number of properties on the same screen or page

Metalanguage should not invade the description (which is the case with feature structures)

Readability of description

Drawbacks
Readable formats are usually not directly exploitable in computer applications
Compilation processes are required

Cf. source code vs. executable code
lemma lexicon vs. inflected-form lexicon

Time limitation

The description of a lexical item is normally limited to a few minutes

Regularities --> classification --> similar items are described in sequence --> efficiency

For properties, description by property is more efficient than description by entry

Even so, manual description of all idioms of a language takes several years

Formal criteria

Formal criteria based on acceptability of sentences

Example: co-referent of possessives
Leurs hôtes préviennent leurs désirs
Leurs hôtes préviennent nos désirs
Leurs hôtes reprennent leurs esprits
**Leurs hôtes reprennent nos esprits*
Identifying such a constraint is immediate for a linguist trained to distributional analysis

Formal criteria

Complementarity between idiom resource and grammar is obtained through distributional analysis

Luc a couché par écrit ses instructions
Luc a <u>mis</u> par écrit ses instructions
**Luc a <u>placé</u> par écrit ses instructions*
**Luc a couché par <u>imprimé</u> ses instructions*
Luc a couché par écrit ses <u>demandes</u>
?**Ses instructions <u>sont</u> par écrit*
--> 2 expressions:
N_0 *coucher par écrit* N_2 N_0 *mettre par écrit* N_2

Formal criteria

Limits of variation are obtained through systematic tests

Luc met cela par écrit
Cela est mis par écrit par Luc

Luc n'entend pas cela de cette oreille
**Cela n'est pas entendu de cette oreille par Luc*

Differential semantic judgment

Comparison of variants
Distributional analysis
Taking into account connotations, implications

Recognising the limits of language plays
Intuition (cf. acceptability judgment, lexicalization, institutionalization)
Requires a corpus also (context: headlines, advertisement...)

Results

Theoretical results
(M. Gross 1982, P. Freckleton 1985, P. Machonis 1985)
'Free' grammar vs. idiom grammar:
Idiom grammar accounts for variants
Idiom grammar is close to free grammar: same structures, same transformations
Idiom entries are more numerous than simple entries

Support verb constructions vs. idioms

Results

Idioms with free determiner, including indefinite determiner

(1)*La défense a <u>cité</u> (un + ce) <u>témoin</u>*
Ce numéro a été <u>le clou de</u> (un + le) <u>spectacle</u> de 2002

Distributionally frozen
The noun with the free determiner can have co-referents, even without language plays

More in core than in periphery

The noun has to be attached both to a simple entry and to the idiom entry
1700 examples like (1), mostly technical

Conclusion

Different backgrounds, different approaches
Backgrounds are so different that much synergy between researchers is missed

A result can have very different users
Theoretical
Practical:
- computer applications
- construction of further resources

Distinct approaches can converge to an objective

Conclusion

Synergy between corpus approaches and introspective approaches

Introspective approaches produce dense, informative resources

Resources are useful to corpus exploration

Corpus exploration is an aid to introspective approaches

Excessive methodological provisions

Throwing away the baby of idiom description with the bath water of introspective linguistics

Bibliographical references

Freckleton, Peter. 1985. Sentence idioms in English, *Working Papers in Linguistics*, University of Melbourne, pp. 153-168 + appendix (196 p.).
Gross, Maurice. 1982. Une classification des phrases "figées" du français, *Revue Québécoise de Linguistique* 11.2, pp. 151-185, Montréal: UQAM.
Machonis, Peter A. 1985. Transformations of verb phrase idioms: passivization, particle movement, dative shift, *American Speech* 60:4, pp. 291-308.
Sastre Martinez, Javier M. 2006. Computer Tools for the Management of Lexicon-Grammar Databases, poster, *Proceedings of the 13th Conference on natural language processing*, TALN 2006, Leuven, 10-13 April 2006, UCL, Presses Universitaires de Louvain, pp. 600-608.