Identification de constructions grammaticales en corpus

Le cas des Constructions Argumentales

Florent Perek – Universität Freiburg & UMR 8163 STL

florent.perek@gmail.com

CerLiCO 23 – 5 Juin 2009

Aperçu

- Approches basées sur corpus dans le cadre des grammaires de constructions
  - Conception de techniques statistiques pour identifier des constructions
  - Problèmes rencontrés
    - Pourquoi ?
    - Implications pour la théorie

Cadre théorique

- La Linguistique de l'usage
  - “usage-based model” (Langacker 1987)
  - Principe de base : l'usage génère la grammaire
    - “grammar is the cognitive organization of one's experience with language” (Bybee 2006:1)
    - L'usage reflète les structures grammaticales
  - Le rôle du corpus est évident
    - comme observatoire de l'usage
    - comme source de données quantitatives

Cadre théorique

- Les Grammaires de Constructions
  - Grammaire = inventaire de signes
  - Pas de séparation entre syntaxe et lexique
- Les Constructions Argumentales (Argument Structure Constructions; Goldberg 1995, 2006)
  - Théorie de la structure argumentale en GC
    - SA = principes gouvernant la réalisation morphosyntaxique des arguments du verbe
    - CA = association d'un sens schématique à des spécifications morphosyntaxiques
    - Indépendantes, non projetées par le verbe

Cadre théorique

- Exemple : la construction ditransitive

p.e. *Mary gave her sister a penny.*

Sam kicked Peter the ball.

Sally baked Sam a cake.

Sem: Agent CAUSES Recipient TO RECEIVE Theme

Syn: $\text{Subject}_{\text{Agent}}$ V $\text{Object1}_{\text{Recipient}}$ $\text{Object2}_{\text{Theme}}$

Cadre théorique

- Problèmes du modèle de Goldberg
  - Importance des particularités lexicales, cf. e.a. Boas (2003) et Iwata (2008)
  - Quel partage des tâches entre entrées lexicales et constructions?
    - Remet en question la généralité des CAs
    - La couverture empirique des CAs reste limitée
      - Une poignée seulement d'exemples récurrents : ditransitive, resultative, caused-motion, intransitive motion, conative et *way*-construction
    - Suggestion: aborder cette question empirique par l'examination de données de corpus

Cadre théorique

- Le corpus en grammaire de constructions
  - Concept de construction problématique pour la linguistique de corpus
    - Annotations centrées sur le mot
    - Pas de moyen direct d'accès au sens
  - Nécessite une analyse manuelle en constructions : chronophage et potentiellement biaisée
  - Paradoxe : ces modèles basés sur l'usage sont rarement documentés par l'usage
  - Besoin de techniques d'exploration de corpus adaptés aux grammaires de construction

Une première étude

- Une première étude : Perek (2008)
  - Objectif : proposer des techniques d'identification des CAs en corpus
  - Nombreuses recherches existantes en acquisition des structures argumentales (surtout en TAL)
    - Acquisition des cadres de sous-catégorisation pour chaque verbe
    - Ici problème différent : pour une forme donnée, quelles sont les CAs ?
  - Restreinte au cas des constructions de la forme :
    - SN V SP
    - SN V SN SP

Une première étude

- Approche statistique
  - Exploite des données formelles et distributionnelles
    - Basée sur le modèle de Goldberg (1995)
    - Annotations syntaxiques : input “propre”
    - Pas de ressources externes, donc pas de sémantique
  - Trois indices statistiques conçus et testés
    - Identification manuelle de trois CAs (dans ICE-GB oral)
    - Test des prédictions statistiques sur les échantillons

Constructions étudiées

- Caused motion construction

  Sem: Agent CAUSES Patient TO MOVE(Path)

  Syn: $\text{Subject}_{\text{Agent}}$ V $\text{Object}_{\text{Patient}}$ $\text{Oblique}_{\text{Path}}$

  p.e. *John threw the ball to the other player.*
  John sneezed the napkin off the table.

- Intransitive motion construction

  Sem: Agent MOVES(Path)

  Syn: $\text{Subject}_{\text{Agent}}$ V $\text{Oblique}_{\text{Path}}$

  p.e. *The ball bounced across the field.*
  The truck rumbled through the tunnel.

Résultats de l'étude

- Un échec relatif
  - Tendance positive mais résultats inutilisables pour l'objectif initial
  - Décalage entre les prédictions du modèle et l'usage réel dans le corpus
  - Approche statistique et formelle trop superficielle : besoin du sens (notamment des verbes)
  - Mais cette étude est instructive car elle nous amène à nuancer la théorie
    - Trois cas pertinents

Problèmes et conséquences (1)

1. Syntaxe créative et augmentation de valence
   - CA capables de contribuer des arguments
     - Rôles constructionnels vs. rôles verbaux
       - p.e. le rôle Path de Caused Motion
         John kicked the ball to the other player.
     - Certains verbes entrent dans une alternance entre
       - une syntaxe simple, c.à.d. avec la valence “naturelle” du verbe
       - une syntaxe étendue = syntaxe simple + rôle constructionnel
     - Dans notre cas :
       - Ditransitive : [Sujet V Objet] vs. [Sujet V Objet1 Objet2]
       - Caused motion : [Sujet V Objet] vs. [Sujet V Objet Oblique]
       - Intransitive motion : [Sujet V] vs. [Sujet V Oblique]

Problèmes et conséquences (1)

- Hypothèses :
  - Certains verbes apparaissent plus fréquemment dans la syntaxe simple que dans la syn. étendue
  - L'existence de tels verbes est signe d'une CA
- Indice basé sur “distinctive collexeme analysis”
  - cf. Stefanowitsch and Gries (2004)
  - Mesure la préférence d'un lexème pour une construction plutôt qu'une autre
  - Ici, entre la syntaxe réduite et la syntaxe étendue

Verbes	Ditransitive (syntaxe simple : Subj-V-Obj)		Caused Motion (syntaxe simple : Subj-V-Obj)		Intransitive Motion (syntaxe simple : Subj-V-Obj)	
Attraction significative pour la syntaxe simple, c.à.d. il y a moins de 5% de chances que la préférence observée pour la syntaxe simple est due au hasard	get	4,51	have	18,37		
	do	4,29				
Pas d'attraction significative pour une syntaxe en particulier, il y a plus de 5% de chances que la préférence observée est due au hasard	leave	-0,47	get	0,52	disappear	-0,3
	save	-0,89	pay	-0,76	sit	-0,31
			set	-0,87	drop	-0,6
			sell	-1,12	pass	-0,6
					close	-0,7
					jump	-0,82
					escape	-1,1
					appear	-1,17
					gather	-1,22
Attraction significative pour la syntaxe étendue (= la construction), c.à.d. il y a moins de 5% de chances que la préférence observée pour la syntaxe étendue est due au hasard	buy	-1,31	**knock**	-1,68	swing	-1,58
	set	-1,49	lend	-1,9	bump	-1,58
	cook	-1,63	**hide**	-2,21	creep	-1,86
	earn	-2,33	**plant**	-2,21	advance	-1,86
	allow	-2,37	offer	-2,22	slip	-1,86
	teach	-3,65	add	-2,31	flow	-2,02
	ask	-4,15	pop	-2,42	descend	-2,32
	send	-7,07	hand	-2,63	stream	-2,32
	offer	-15,18	refer	-2,71	fall	-2,33
	show	-20,69	**shove**	-3,18	**burst**	-3,16
	tell	-57,09	post	-3,18	tread	-3,48
	give	-191,87	sit	-4,18	enter	-3,48
			leave	-4,31	fly	-3,52
			impose	-5,21	spread	-3,8
			take	-5,33	walk	-4,22
			bring	-5,48	move	-4,25
			throw	-6,4	head	-4,42
			place	-8,14	run	-6,2
			send	-13,26	get	-7,31
			give	-25,41	embark	-8,12
			put	-63,16	return	-8,85
					come	-32,79
					go	-41,21

Problèmes et conséquences (1)

- Deux écueils
  - Syntaxe créative rare en corpus
  - Les rares cas sont plus fréquents en syn. étendue
- Conséquences théoriques
  - canonique ≠ fréquent
  - Décalage entre théorie et usage
    - En linguistique de l'usage, fréquence → enracinement (*entrenchment*)
    - Pourquoi la syntaxe étendue n'est-elle pas perçue comme la valence naturelle du verbe ?
    - Statut de la dichotomie entrée lexicale / construction ?

Problèmes et conséquences (2)

2.Polysémie des prépositions

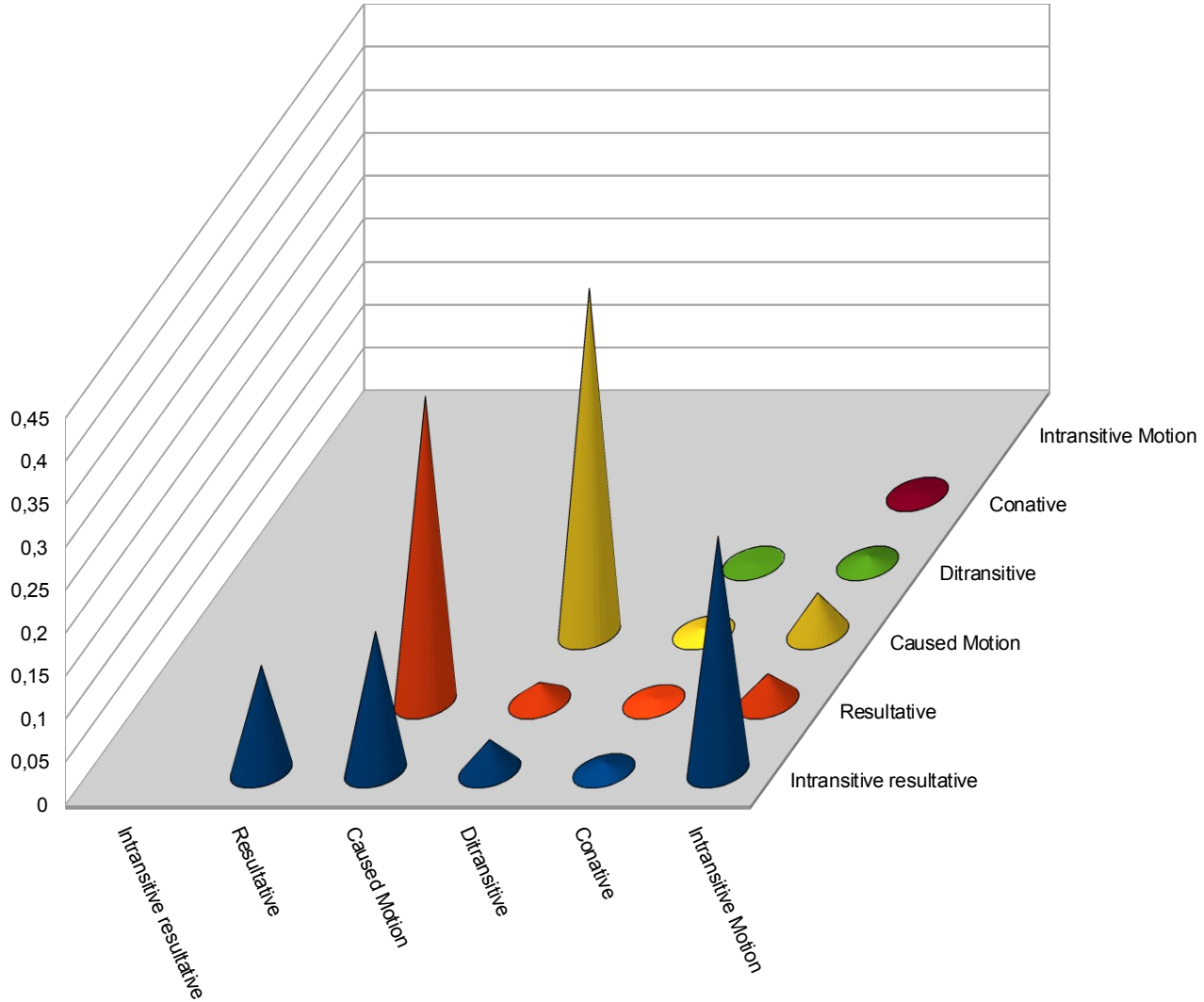
- SN V *at* SN : intransitive motion vs. conative
- SN V *to* SN : intransitive motion vs. *speak, happen, listen, amount…*
- L'interprétation dépend du verbe
- La préposition est-elle un niveau de description formelle adéquat ?
  - Dépend des constructions : parfois imposée par la syntaxe, parfois choisie pour sa contribution sémantique
  - Identification d'une construction = catégorisation; quel conséquence pour le pôle syntaxique ?

Problèmes et conséquences (3)

3.La variation du sens constructionnel

- Postulat : la distribution verbale reflète le sens constructionnel, cf. :
  - Origine lexicale du sens constructionnel
  - Principe de compatibilité sémantique : la construction attire des verbes ayant un certain sens
- Comparaison sémantique des constructions par la distribution verbale
  - Résultat fidèle à l'intuition
  - Distinctions plus fines par les arguments ?

Problèmes et conséquences (3)



Problèmes et conséquences (3)

- Suggestion : utiliser cette technique pour unifier des variantes formelles
  - c.à.d. des utilisations de la construction avec des prépositions différentes
  - Fonctionne dans certains cas, p.e. NP V *to* NP et NP *into* NP ont des distributions similaires
  - Mais dans d'autres cas, la distribution est très différente, p.e. *around* et *under*
  - La polysémie pose ici aussi problème
- Conséquence pour la théorie : peut-on parler d'une seule construction ?

Conclusion

- Des leçons tirées de cette expérience
  - Une méthode purement formelle est vaine : nécessité de réintégrer la sémantique
    - Mais comment : dictionnaire (e.g. Wordnet), analyse distributionnelle (Latent Semantic Analysis), … ?
    - Sémantique des verbes, prépositions, ... ?
    - Rôle des arguments ?
  - cf. Alishahi & Stevenson (2008) : apprentissage statistique guidé par annotations sémantiques

Conclusion

- Suggère des ajustements pour la théorie
  - Dans la relation entre usage et structure
  - Dans la définition de la forme syntaxique
  - Pose la question du statut des constructions en tant qu'unités discrètes

Bibliographie

Alishahi, A. et S. Stevenson. 2008. A computational model for early argument structure acquisition. *Cognitive Science* 32, 5, 789-834.

Boas, H. C. (2003). *A Constructional Approach to Resultatives*. CSLI, Stanford.

Bybee, J. (2006). From usage to grammar : The mind's response to repetition. *Language*, 82(4):711–733.

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language.* Oxford University Press, Oxford.

Gries, S. Th. & A. Stefanowitsch (2004). Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9:1, 97-129.

Iwata, S. (2008). *Locative Alternation: A lexical-constructional approach*. John Benjamins, Amsterdam.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar. Vol. I.* Stanford University Press, Stanford.

Perek, F. (2008). Towards a constructional approach to automatic argument structure acquisition: the case of oblique phrases. Mémoire de Master, Université Charles de Gaulle Lille III, Villeneuve d'Ascq, France.