Passive-regressive: grammatical voice alternations modelling and feature analysis

Liubov Nesterenko
National Research University Higher School of Economics, Moscow
lnesterenko@hse.ru
22 November 2018

Introduction

Voice phenomenon

(1) a. John bought this house for $250000 in 1980.
b. This house was bought by John for $250000 in 1980.

Voice phenomenon

(1) a. John bought this house for $250000 in 1980.
b. This house was bought by John for $250000 in 1980.

Active voice vs. passive voice

The situation is same but the language encoding is different.

Voice function

Voice function is mapping the semantic structure of the verb (semantic roles) to the syntactic structure.

(1) a. John bought this house for $250000 in 1980.

The agent is a (syntactic) subject, the most prominent constituent and the patient is an object, a less prominent constituent.

b. This house was bought by John for $250000 in 1980.

The patient becomes topical and the agent is demoted.

Semantic roles

Semantic roles

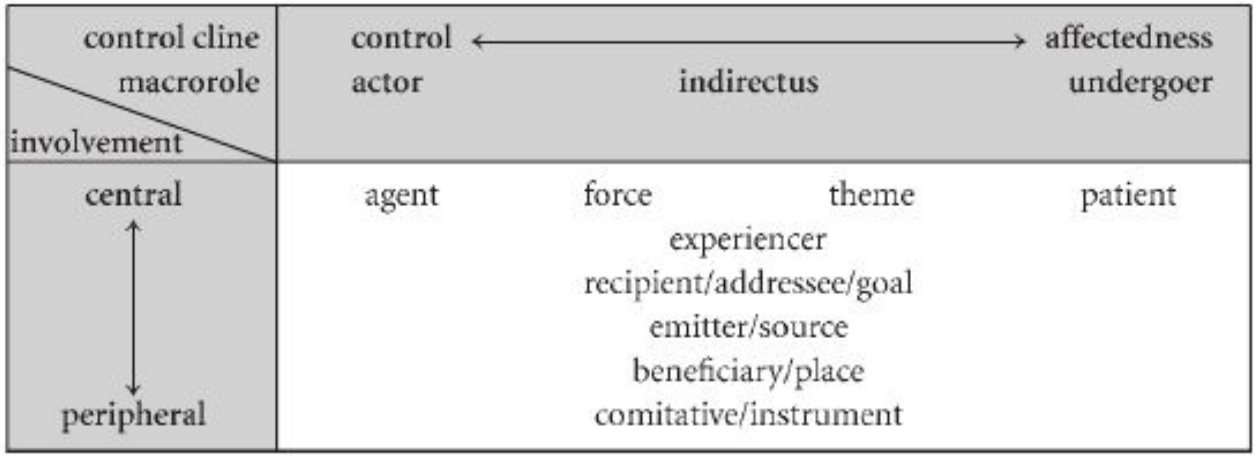
(2) Agent & Patient
Linda *solved the problem.*

(3) Agent & Experiencer
He *impressed me.*

(4) Agent & Theme
I *found my keys.*

(5) Experiencer & Theme
I *saw Molly.*

Macroroles, involvement and control



(Lehmann, 2006)

- The participant that has **most control** in the situation is the **actor**, the one that is **most controlled** is the **undergoer**.
- A participant is **maximally involved** in the situation if the situation is inconceivable without this participant.

Goal of the study

The goal of the study is to determine *which contextual and semantic factors influence the choice* of a voice construction cross-linguistically.

Methods

Language, models and predictions

Traditional methods

Analysis of samples from grammars or from corpora in order to formulate rules that motivate the choice between different means of expression for particular linguistic phenomenon.

Language, models and predictions

Traditional methods

Analysis of samples from grammars or from corpora in order to formulate rules that motivate the choice between different means of expression for particular linguistic phenomenon.

Alternative methods

Logistic regression modelling.

Use of parallel corpora and features transfer.

Analysis of features that were used for model training.

Language, models and predictions

This work is inspired by the study of (Bresnan, Ford 2010).

Language, models and predictions

This work is inspired by the study of (Bresnan, Ford 2010).

Dative alternation

- Double object construction
  showed the woman the ticket
- Dative with “to”
  showed the ticket to the woman

Language, models and predictions

This work is inspired by the study of (Bresnan, Ford 2010).

Dative alternation

- Double object construction
  showed the woman the ticket
- Dative with “to”
  showed the ticket to the woman

Linguistic features

Pronominality of recipient = pronoun
Pronominality of theme = pronoun
Definiteness of recipient = indefinite
Definiteness of theme = indefinite
Animacy of recipient = inanimate
Number of theme = plural
...
etc.

Data

- A corpus of Harry Potter books in 9 languages
  English, German, Swedish, French, Spanish, Italian, Russian, Czech, Bulgarian

Data

- A corpus of Harry Potter books in 9 languages
  English, German, Swedish, French, Spanish, Italian, Russian, Czech, Bulgarian
- ~ 1 million tokens per language

Data

- A corpus of Harry Potter books in 9 languages
  English, German, Swedish, French, Spanish, Italian, Russian, Czech, Bulgarian
- ~ 1 million tokens per language
- Texts were parsed with UDPipe parser

Data

- A corpus of Harry Potter books in 9 languages
  English, German, Swedish, French, Spanish, Italian, Russian, Czech, Bulgarian
- ~ 1 million tokens per language
- Texts were parsed with UDPipe parser
- Passive sentences were automatically extracted (only passives with overtly expressed agent)

Data

- A corpus of Harry Potter books in 9 languages
  English, German, Swedish, French, Spanish, Italian, Russian, Czech, Bulgarian
- ~ 1 million tokens per language
- Texts were parsed with UDPipe parser
- Passive sentences were automatically extracted (only passives with overtly expressed agent)
- Since it's (in general) impossible for an intransitive verb to build a passive sentence with an overtly expressed agent, only active sentences with transitive predicates were taken into consideration.

The general scheme

- Determining relevant linguistic features

The general scheme

- ➢ Determining relevant linguistic features
- ➢ Manually annotating the English data according to the feature set

The general scheme

- Determining relevant linguistic features
- Manually annotating the English data according to the feature set
- Projecting these features to the parallel data with active/passive labels from other languages

The general scheme

- ➢ Determining relevant linguistic features
- ➢ Manually annotating the English data according to the feature set
- ➢ Projecting these features to the parallel data but using active/passive labels from other languages
- ➢ Training the models for active/passive prediction

The general scheme

- Determining relevant linguistic features
- Manually annotating the English data according to the feature set
- Projecting these features to the parallel data but using active/passive labels from other languages
- Training the models for active/passive prediction
- Analysing feature weights, comparing and evaluating the models

The general scheme

- ➢ Determining relevant linguistic features
- ➢ Manually annotating the English data according to the feature set
- ➢ Projecting these features to the parallel data but using active/passive labels from other languages
- ➢ Training the models for active/passive prediction
- ➢ Analysing feature weights, comparing and evaluating the models

Important!

It is not the task of syntactic parsing.

The use of surface features is restricted

Model: features

AS - active subject, PS - (potential) passive subject

Linda$_{AS}$ *solved the problem*$_{PS}$

Model: features

AS - active subject, PS - (potential) passive subject

Linda$_{\textbf{AS}}$ *solved the problem*$_{\textbf{PS}}$

Contextual features

- AS has referent
- PS has referent

  *E.g. When **Harry** pulled back his sheets, **he** found his Invisibility Cloak folded neatly underneath them.*

Model: features

AS - active subject, PS - (potential) passive subject

Linda$_{AS}$ *solved the problem*$_{PS}$

Contextual features

- AS has referent
- PS has referent

  *E.g. When **Harry** pulled back his sheets, **he** found his Invisibility Cloak folded neatly underneath them.*

- Contrast

  *E.g. Harry tried to argue back **but** his words were drowned by a long, loud belch from the Dursleys' son, Dudley.*

Model: features

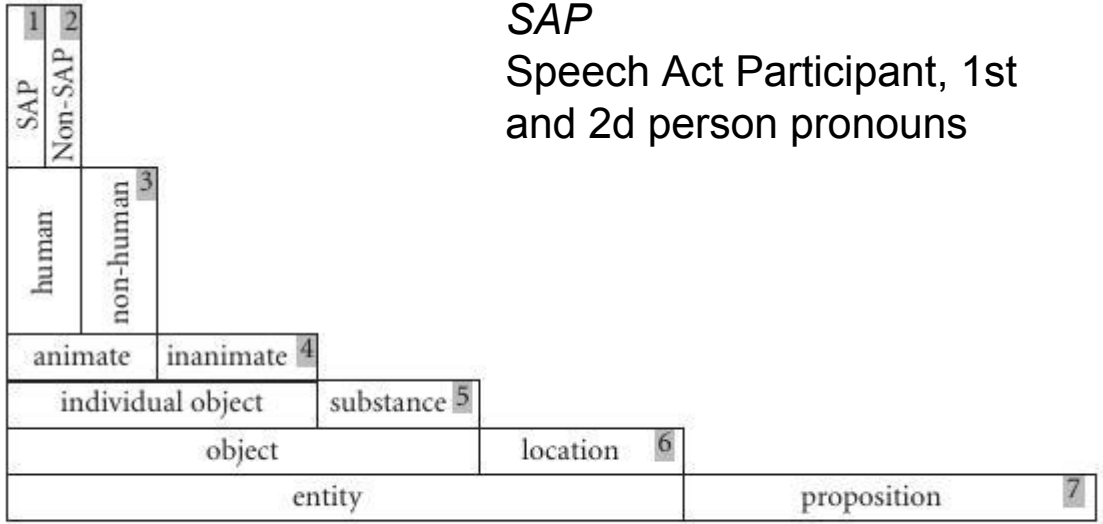
Semantic features

- AS is actor
- PS is undergoer
- Involvement of AS
- Involvement of AS
- Empathy of AS
- Empathy of PS
- Empathy difference
- Empathy difference_2

Model: features

Semantic features

- AS is actor
- PS is undergoer
- Involvement of AS
- Involvement of AS
- Empathy of AS
- Empathy of PS
- Empathy difference
- Empathy difference_2



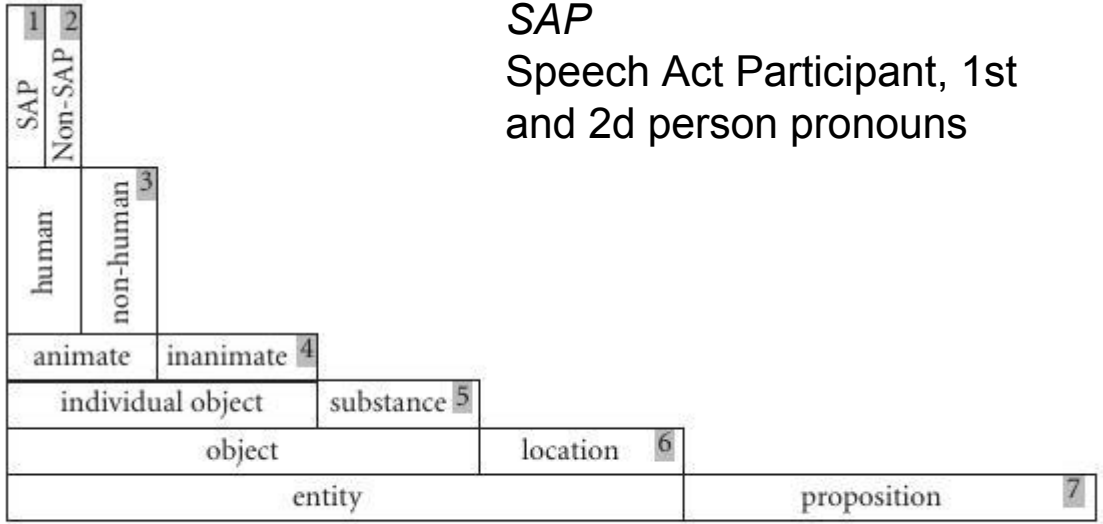
SAP
Speech Act Participant, 1st and 2d person pronouns

(Lehmann, 2006)

Model: features

Semantic features

- AS is actor
- PS is undergoer
- Involvement of AS
- Involvement of AS
- Empathy of AS
- Empathy of PS
- Empathy difference
- Empathy difference_2



SAP
Speech Act Participant, 1st and 2d person pronouns

(Lehmann, 2006)

(6) a. $250000 won't by this kind of house.
b. ??This kind of house won't be bought by $250000. (Shibatani, 1985)

Results

English model 1

Feature	Weight	p-value
AS is actor	2.5434	0.002
Empathy of AS	1.8749	0.000
AS has referent	-3.3721	0.000
Empathy of PS	-0.4811	0.024
PS has referent	1.4064	0.001
Involvement of PS	7.3540	0.004
Empathy distance	2.5812	0.000
Empathy distance 2	-2.3021	0.005
Contrast	0.2884	0.020
Intercept	-11.7233	0.000

Set	Total	Active/Passive
Train	307	0.49/0.51
Test	43	0.45/0.55
Accuracy	0.86	

English, Swedish, German

Feature	English Weight	p-value	Swedish Weight	p-value	German Weight	p-value
AS is actor	1.9966	0.021	1.0432	0.276	2.3059	0.012
Empathy of AS	1.6693	0.000	0.7365	0.035	0.9262	0.003
AS has referent	-3.4933	0.000	-2.7681	0.010	-2.2080	0.004
Empathy of PS	-0.5675	0.061	0.1665	0.553	-0.3124	0.180
PS has referent	1.4786	0.004	1.2022	0.048	0.8925	0.069
Involv. of PS	4.0700	0.157	6.2070	0.075	3.4101	0.204
Contrast	2.1334	0.002	2.0539	0.001	0.6384	0.222
Empathy dist	-1.9692	0.050	-2.8253	0.022	-0.3378	0.731
Empathy dist2	0.2357	0.152	0.2773	0.131	-0.0054	0.973
Intercept	-8.3810	0.001	-10.4443	0.002	-7.8513	0.003

	Train A/P	Test A/P	
English	Train 0.68 0.32	Test 0.66 0.34	Accuracy 0.87
Swedish	Train 0.84 0.16	Test 0.81 0.19	Accuracy 0.87
German	Train 0.83 0.17	Test 1.0 0	Accuracy 1.0
Total	Train 218 (0.87)	Test 32 (0.13)	

English, French, Italian, Spanish

	English		French		Italian		Spanish	
Features	**Weight**	**p-value**	**Weight**	**p-value**	**Weight**	**p-value**	**Weight**	**p-value**
AS is actor	1.9966	0.021	1.2856	0.107	2.4192	0.022	0.1558	0.894
Empathy of AS	1.6693	0.000	0.6807	0.017	1.1388	0.003	0.2158	0.672
AS has referent	-3.4933	0.000	-3.6624	0.001	-2.9115	0.000	-12.2958	0.954
Empathy of PS	-0.5675	0.061	-0.8081	0.003	-0.1182	0.648	-0.8390	0.065
PS has referent	1.4786	0.004	0.7573	0.131	0.8960	0.069	0.8050	0.356
Involv. of PS	4.0700	0.157	6.2599	0.025	6.9770	0.020	9.1327	0.145
Contrast	2.1334	0.002	2.5458	0.000	1.6477	0.006	-1.2284	0.286
Empathy dist	-1.9692	0.050	-0.6100	0.544	-2.1054	0.036	-2.5028	0.173
Empathy dist 2	0.2357	0.152	0.0529	0.745	0.2055	0.191	0.3572	0.233
Intercept	-8.3810	0.001	-7.8390	0.002	-11.4761	0.000	-9.9772	0.107

	Train A/P	Test A/P	
English	Train 0.68 0.32	Test 0.66 0.34	Accuracy 0.87
French	Train 0.79 0.21	Test 0.96 0.04	Accuracy 0.9
Italian	Train 0.76 0.24	Test 0.93 0.07	Accuracy 0.84
Spanish	Train 0.94 0.6	Test 0.9 0.1	Accuracy 0.9
Total	Train 218 (0.87)	Test 32 (0.13)	

English, Russian, Czech, Bulgarian

	English		Russian		Czech		Bulgarian	
Features	**Weight**	**p-value**	**Weight**	**p-value**	**Weight**	**p-value**	**Weight**	**p-value**
AS is actor	1.9966	0.021	1.4709	0.115	1.2673	0.205	1.4686	0.085
Empathy of AS	1.6693	0.000	1.2416	0.000	0.7114	0.059	1.2004	0.001
AS has referent	-3.4933	0.000	-1.1870	0.152	-8.5146	0.748	-8.8394	0.583
Empathy of PS	-0.5675	0.061	-0.0577	0.837	-0.1335	0.671	-0.4124	0.087
PS has referent	1.4786	0.004	0.0912	0.885	0.6102	0.345	0.8298	0.088
Involv. of PS	4.0700	0.157	1.0955	0.739	3.4414	0.348	1.4490	0.576
Contrast	2.1334	0.002	-0.8321	0.286	0.6707	0.303	0.5919	0.278
Empathy dist	-1.9692	0.050	-1.6904	0.180	-0.8649	0.533	-1.0220	0.263
Empathy dist 2	0.2357	0.152	0.1858	0.318	0.0048	0.983	0.1431	0.318
Intercept	-8.3810	0.001	-6.7771	0.043	-7.3485	0.035	-5.1970	0.031

	Train A/P	Test A/P	
English	Train 0.68 0.32	Test 0.66 0.34	Accuracy 0.87
Russian	Train 0.9 0.1	Test 0.85 0.15	Accuracy 0.84
Czech	Train 0.92 0.08	Test 0.97 0.03	Accuracy 0.96
Bulgarian	Train 0.8 0.2	Test 0.71 0.29	Accuracy 0.78
Total	Train 218 (0.87)	Test 32 (0.13)	

Features summary (based on the English model)

Passive-pro features

- contrast
- involvement of PS
- PS has referent
- AS is actor
- empathy of AS
- empathy distance 2

Active-pro features

- empathy distance
- AS has referent
- empathy of PS

Conclusion

- More data needed for training the models
- Correcting mistakes of UDPipe active/passive might help to improve the results
- As for the features, the overall tendency for active-pro passive-pro features is similar across languages, but a deeper insight is needed.
- Error analysis might reveal some language-specific information about voice constructions