Syntactic Theory and Machine Learning: focus on German and German varieties

Giuseppe Samo

Department of Linguistics,

Beijing Language and Culture University

北京语言大学语言学系

samo@blcu.edu.cn

北京语言大学语言学系
DEPARTMENT OF LINGUISTICS

Roadmap

- *Syntactic theory & frequency*
- *A set of research questions*
- *A case study: testing theoretical proposals*
  - *Some notes on the syntax of German*
  - *Materials & Methods*
  - *Results and discussion*
- *Further improvements*

Take home message

- *Observational data can be adopted as a measure to test a subset of linguistic proposals in formal theory;*
- *An important role is played by the ability of “translating” theories;*
- *Syntactic theory should make use, when possible, of statistical measures to:*
    - *Compare theoretical proposals*
    - *Create new research questions*

Syntactic Theory and observational data

- *Usage shapes Grammar* (Ibbotson 2013 for an overview)
- *Colorless green ideas sleep furiously* (but see also Bresnan et al. 2001; Yang 2017, 2018, see also Yang et al. 2017 in acquisition, etc.)
- Frequency as a dependent variable to test linguistic proposals (Merlo 1994; Merlo & Stevenson 1998; Merlo 2015, 2016; Samo & Merlo 2019).

Machine Learning and Syntactic Theory

- (Naïve Bayes) classifiers and syntactic theory
  - Test the classification ability of (and compare) proposal (Merlo 2015)
  - Find the “weight” of syntactic operations (Merlo & Ouwayda 2018)
  - Predict classes (e.g., Zimmermann 2014, “dating manuscripts on the basis of text-internal criteria in language change environments)

Machine Learning and Syntactic Theory

- *Deep learning and syntactic structure*
    - A state of the art in Linzen & Baroni (2021)
    - Machines are able to perform complex, e.g. long distance agreement also with non sensical sentences (Gulordava et al. 2018 *inter alia*)
    - "Black box"; Lakretz et al. (2019)

Test linguistic proposals

Merlo (2015)

Universal 20: Dem Num Adj N

Three theories: Cinque (2005), Cysouw (2010) & Dryer (2006)

Merlo & Ouwayda (2018)

Dem Num Adj N vs. Dem Adj Num N

A set of research questions when working on one language (e.g. German)

1. Can we test fine-grained elements of theoretical proposals *[e.g. Cartography of Syntactic Structure, Cinque & Rizzi 2010, Rizzi & Cinque 2016]* ?
2. Can we detect dimensions of micro-variation according to the nature of the treebank in terms of *genres/registers* *[e.g., Samo et al. 2020]?*
3. Can we, when dataset are available, observe micro-variation among varieties?

A study on Word Order in German

- *Some notes on the syntax of German*
- *Materials & Methods*
- *Results and discussion*

Samo, G. (2019b) Cartography and Locality in German: a quantitative study with Dependency structures , Rivista di Grammatica Generativa/Research in Generative Grammar, 5, 1-26, ISSN 2531-5935.

https://lingbuzz.com/j/rgg/2019/2019.05/samo_cartography-and-locality-in-german_RGG-2019-05.pdf

Verb Second (V2)

(1) a. *Die Katze hat gestern den Apfel gegessen*

The cat has yesterday the apple eaten

‘The cat ate the apple yesterday’

b. *Den Apfel hat die Katze gestern gegessen.*

c. *Gestern hat die Katze den Apfel gegessen.*

Around the verb

Restrictions and freedom of movement of syntactic *reorderings* in (West) German(ic):

Vorfeld (‘prefield): locus of limited movement;

Mittelfeld (‘middlefield’): locus of extreme flexibility.

Vorfeld

The layer of the structure preceding (syntactically higher than) the inflected verb (*Vorfeld* 'prefield') **seems** inaccessible to more than one constituent.

Violations to V2: 🕮 **West Germanic** Standard German (S. Müller 2013; Meinunger 2004), Kietzdeutsch (Wiese 2009; Walkden 29017), West Flemish and Standard Dutch (De Clercq & Haegeman 2018, Haegeman & Greco 2018, *inter alia*). 🕮 **Scandinavian** Norwegian (Nilsen 2003, Wiklund et al. 2007), Tromsø-Norwegian (Westergaard, Øystein, Lohndal 2012), Swedish (Bohnacker 2006), Icelandic (Þrainsson 2007), Urban Vernaculars of Danish, Norwegian and Swedish (Walkden 2017).

Vorfeld: only one constituent

(2) a. *_Den Apfel die Katze hat gestern gegessen._

the apple the cat has yesterday eaten

b. *_Gestern den Apfel hat die Katze gegessen._

c. *_Gestern die Katze hat den Apfel gegessen._

d. *_Die Katze gestern den Apfel hat gegessen._

Mittelfeld: scrambling constituents

On (literally) the other side (of the verb), the portion right after the inflected verb (*Mittelfeld*' Middlefield') is depicted as a locus of <u>extreme flexibility </u>for the movement of syntactic elements (in West Germanic).

Scrambling (Lenerz 1977, Frey 2004, Hinterhölzl 2006, see also Schoenmakers 2020 for Dutch): syntactic constituents seems to be freely placed, as shown in (3), with different degrees of acceptability.

German (adapted from Samo 2019a: 60-62; 30-35)

(3) a. *Die Katze hat DEM HUND den Apfel gegeben.*

The cat.nom has the.dat dog the.acc apple given.

b. Die Katze hat *den Apfel* *DEM HUND* gegeben.

c. *Den Apfel* hat die Katze *DEM HUND* gegeben.

d. *Den Apfel* hat *DEM HUND* die Katze gegeben.

e. *DEM HUND* hat die Katze *den Apfel* gegeben.

f. *DEM HUND* hat *den Apfel* die Katze gegeben.

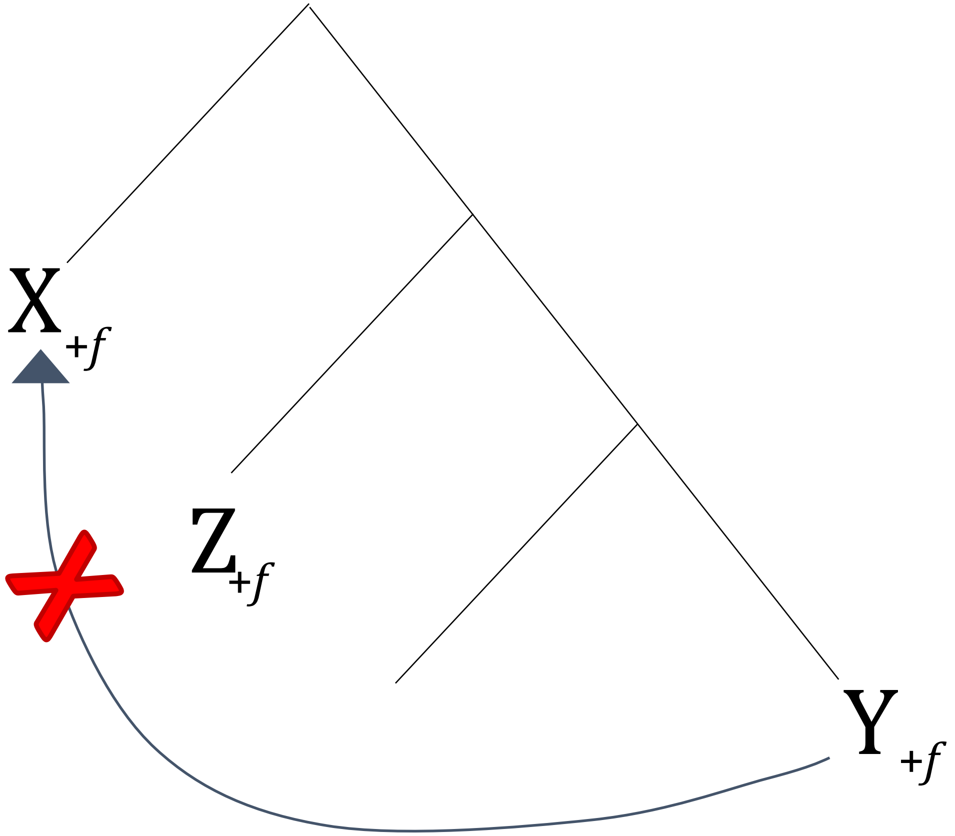
Formal accounts

Different theories (Den Besten 1983, Haegeman 1996, Poletto 2002, Holmberg 2015, Wolfe 2016, Abels 2017, *inter alia*).

Samo (2019a) *"A criterial approach to the Cartography of V2", John Benjamins Publishing*.

V2 and Scrambling reduced ***to* only one phenomenon**. Locality effects in terms of featural relativized minimality (Rizzi 1990, 2004; Starke 2001; Friedmann et al. 2009)

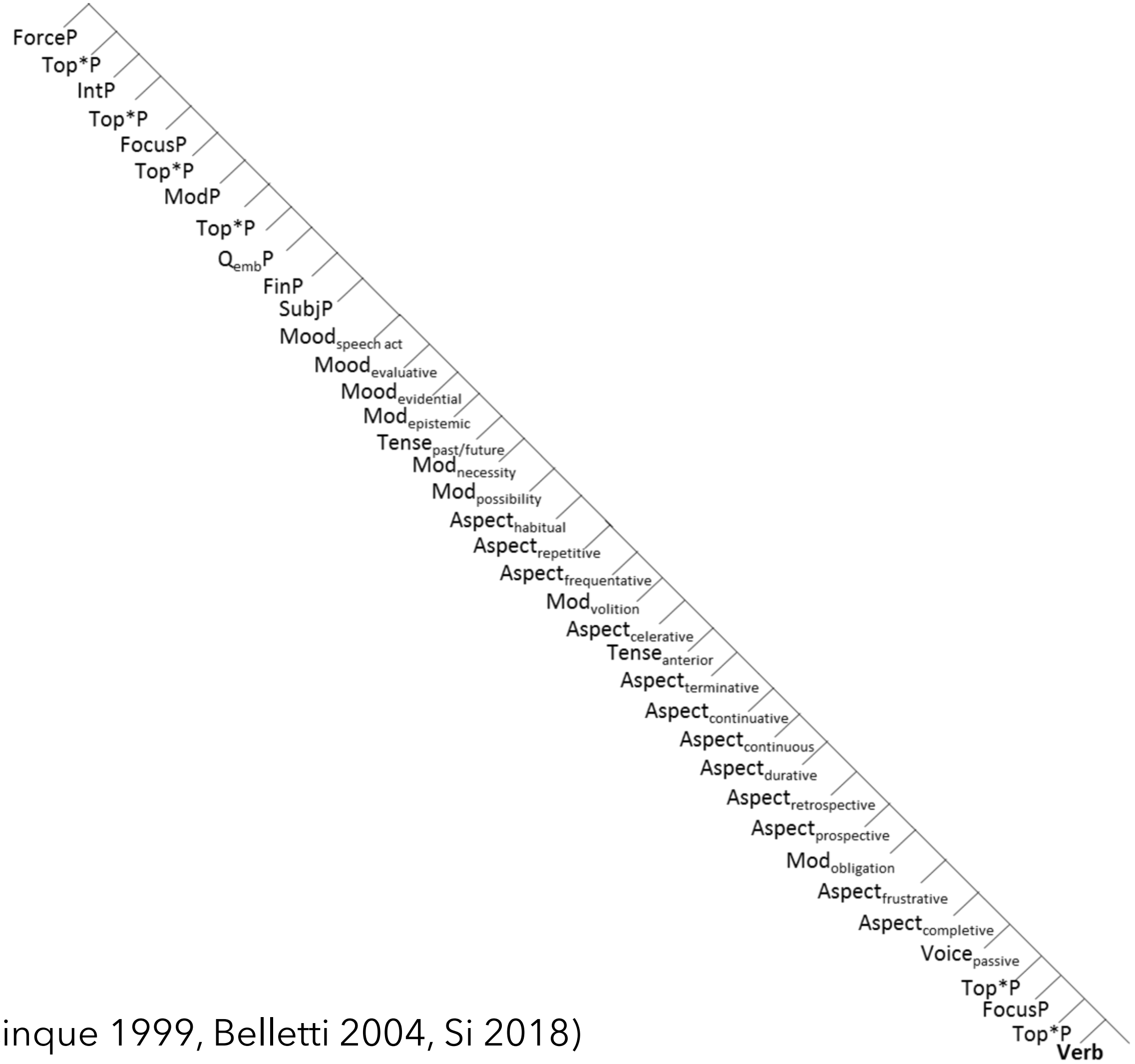
Locality (Starke 2001; Rizzi 1990, 2004)



Y = Generation Site
X = Landing Site
Z = Intervener

+f = classes of features, features

Base and Generation sites with a cartographic approach (Rizzi & Cinque 2016 *inter alia*)


ForceP
Top*P
IntP
Top*P
FocusP
Top*P
ModP
Top*P
Q$_{\text{emb}}$P
FinP
SubjP
Mood$_{\text{speech act}}$
Mood$_{\text{evaluative}}$
Mood$_{\text{evidential}}$
Mod$_{\text{epistemic}}$
Tense$_{\text{past/future}}$
Mod$_{\text{necessity}}$
Mod$_{\text{possibility}}$
Aspect$_{\text{habitual}}$
Aspect$_{\text{repetitive}}$
Aspect$_{\text{frequentative}}$
Mod$_{\text{volition}}$
Aspect$_{\text{celerative}}$
Tense$_{\text{anterior}}$
Aspect$_{\text{terminative}}$
Aspect$_{\text{continuative}}$
Aspect$_{\text{continuous}}$
Aspect$_{\text{durative}}$
Aspect$_{\text{retrospective}}$
Aspect$_{\text{prospective}}$
Mod$_{\text{obligation}}$
Aspect$_{\text{frustrative}}$
Aspect$_{\text{completive}}$
Voice$_{\text{passive}}$
Top*P
FocusP
Top*P
Verb

Maps in (Rizzi 1997, Rizzi & Bocci 2017, Cardinaletti 2004, Cinque 1999, Belletti 2004, Si 2018)

Types of intervention

	X	**Z**	**Y**	**Type of relation**	**Children**[1]	**Adults**
a.	+A	**+A**	<+A>	identity	*	*
b.	+A,+B	**+A**	<+A,+B>	inclusion	*	*harder*
c.	+A,+B	**+A,+C**	<+A,+B>	intersection	ok	ok
d.	+A	**+B**	<+A>	disjunction	ok	ok

(Martini et al. 2018)

[1] Also atypical development (Durlemman et al. 2015) and language pathologies (e.g. Aphasia, Grillo 2008; Martini et al. 2019).

Testing locality in grammatical clauses

- *Syntactically annotated corpora*
- *A translation from dependencies into syntactic functional projections*
- *Frequencies!*

Retrievable in observational studies.
***Grammatical** but harder.*

Grammatical clauses

	X	Z	Y	Type of relation	Children[1]	Adults
a.	+A	**+A**	<+A>	identity	*	*
b.	+A,+B	**+A**	<+A,+B>	inclusion	*	*harder*
c.	+A,+B	**+A,+C**	<+A,+B>	intersection	ok	ok
d.	+A	**+B**	<+A>	disjunction	ok	ok

(Martini et al. 2018)

[1] Also atypical development (Durlemman et al. 2015) and language pathologies (e.g. Aphasia, Grillo 2008; Martini et al. 2019).

Syntactically annotated corpora

- Universal Dependencies (Nivre 2015, Zeman et al. 2020);
- +100 languages, +150 treebanks annotated under the same guidelines;
- Syntactic dependencies, POS and morphological annotations;
- Figure from (Samo & Merlo 2019: 4, Fig.1 and Fig.2)

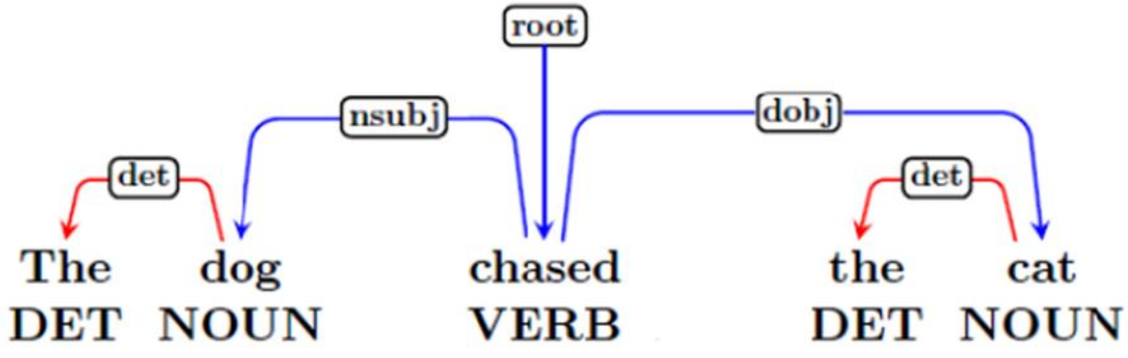


Figure 1: Canonical order

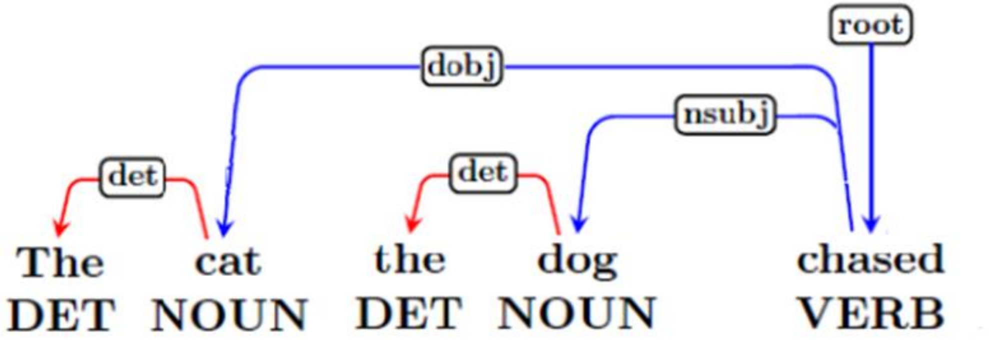
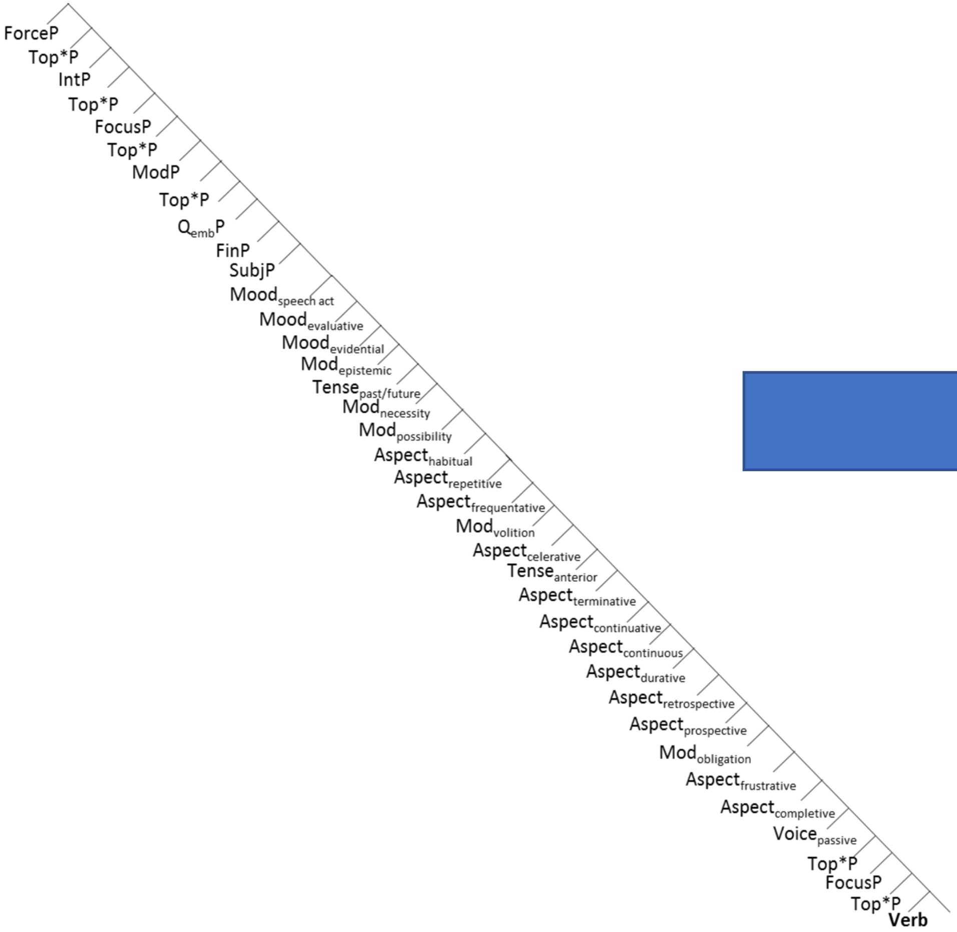


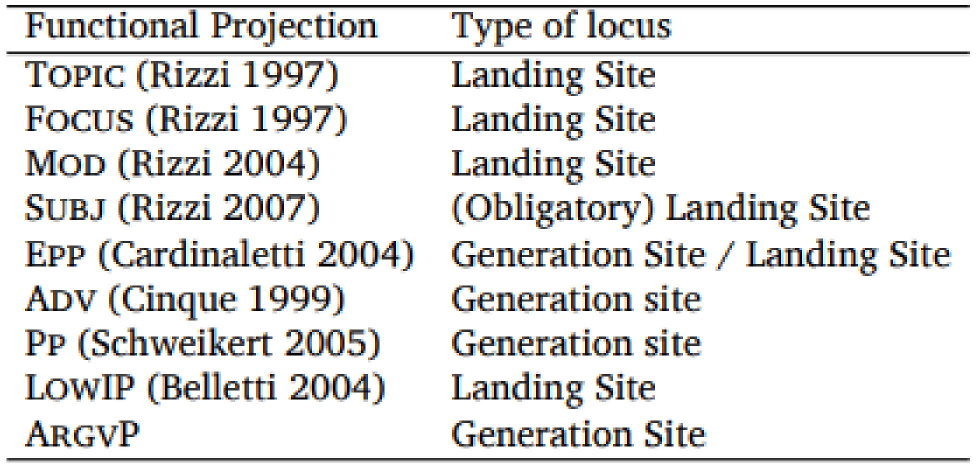
Figure 2: Non-canonical order

Large-corpora annotation (e.g. UD, Nivre 2015)

- NSUBJ: subjects
- OBJ: objects
- IOBJ: indirect objects
- OBL: oblique (complements)
- ADVMOD: adverbial modifiers
- EXPL: expletives (not only subject expletives!)

Translating frameworks (from Samo 2019b)

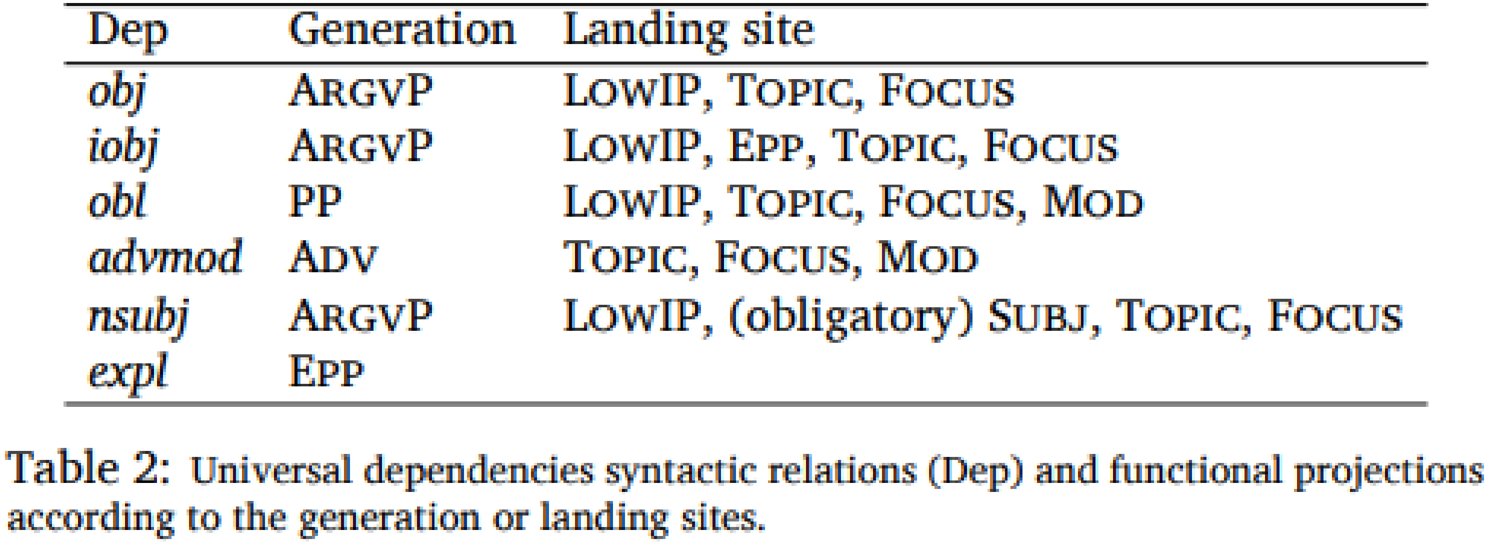




Functional Projection	Type of locus
Topic (Rizzi 1997)	Landing Site
Focus (Rizzi 1997)	Landing Site
Mod (Rizzi 2004)	Landing Site
Subj (Rizzi 2007)	(Obligatory) Landing Site
Epp (Cardinaletti 2004)	Generation Site / Landing Site
Adv (Cinque 1999)	Generation site
Pp (Schweikert 2005)	Generation site
LowIP (Belletti 2004)	Landing Site
ArgvP	Generation Site

Table 1: Nature (generation site or landing site) of functional projections (and related references).

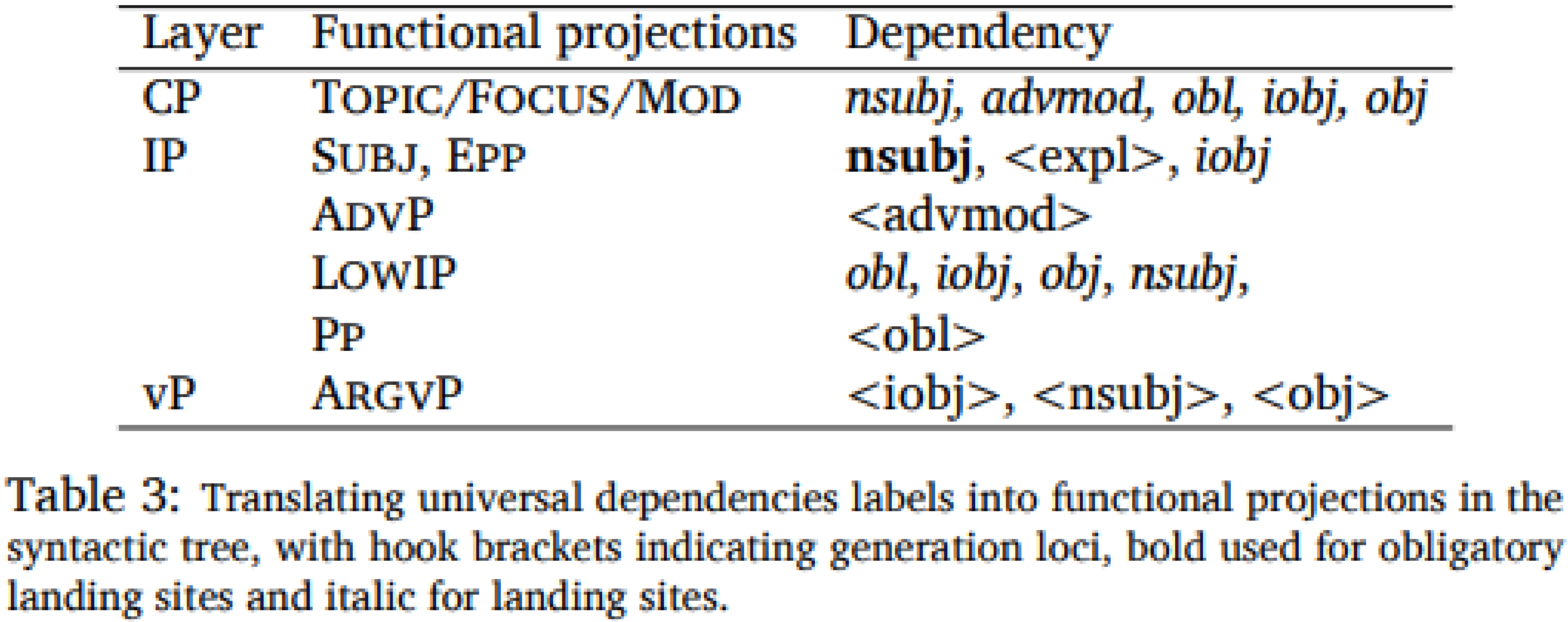
Translating frameworks



Dep	Generation	Landing site
obj	ArgVP	LowIP, Topic, Focus
iobj	ArgVP	LowIP, Epp, Topic, Focus
obl	PP	LowIP, Topic, Focus, Mod
advmod	Adv	Topic, Focus, Mod
nsubj	ArgVP	LowIP, (obligatory) Subj, Topic, Focus
expl	Epp	

Table 2: Universal dependencies syntactic relations (Dep) and functional projections according to the generation or landing sites.

Samo (2019b)



Layer	Functional projections	Dependency
CP	Topic/Focus/Mod	*nsubj, advmod, obl, iobj, obj*
IP	Subj, Epp	**nsubj**, <expl>, *iobj*
	AdvP	<advmod>
	LowIP	*obl*, *iobj*, *obj*, *nsubj*,
	Pp	<obl>
vP	ArgvP	<iobj>, <nsubj>, <obj>

Table 3: Translating universal dependencies labels into functional projections in the syntactic tree, with hook brackets indicating generation loci, bold used for obligatory landing sites and italic for landing sites.

Features triggering locality

- NSUBJ: subjects
- OBJ: objects
- IOBJ: indirect objects
- OBL: oblique (complements)
- ADVMOD: adverbial modifiers

Theory α.
Argumental: *nsubj, obj, iobj*
Non-argumental: advmod, obl

Classical analysis

Theory β.
Argumental: *nsubj, obj, iobj*
Non-argumental: advmod, obl

(based on Schweikert 2005)

Materials

Treebank	Trees	Tokens	Genre
UD_German-HDT@2.7	189928	3589318	News, nonfiction, web
UD_German-GSD@2.7	15590	308387	News, reviews, wiki
UD_German-PUD@2.7	1000	22329	News, wiki
UD_German-LIT@2.7	1922	42362	Nonfiction

Treebanks, size and genres

HDT: Hamburg Dependency Treebank (Borges Völker et al., 2019)

GSD, PUD, LIT: www.universaldependencies.com [relevant treebank pages]

Structures

A – V2 – B

A - V2 – […] – B

V2 – A – B

Tool: Grew-match maintained by Inria in Nancy, http://match.grew.fr/

Naturally occurring examples

Pattern	Example	ID
advmod – V2 – subj	*Hier lasse ich mein Geld gerne!* Here leave 1sg my money gladly	GSD, dev-s337
obj – V2 – […] – obl	*Textmeldungen stellt sie auf einem LC-Display dar* Text-messages shows 3sgf on an LC display	HDT, hdt-s10956
subj > iobj	*[…], die der Präsident ihnen gestern übermitteln wollte.* that the President 3pl.dat yesterday to-convey wanted	PUD, n05003019

Examples of naturally occurring examples extracted from the treebanks (TB) and relative sentence IDs in specific patterns.

Order of constituents (German)

A	B	A-V2-[…] - B	A-V2 - B	A > B	%CE
Subj	Obj	8525	26307	88934	**0.88**
Obj	Subj	2296	2895	12468	0.12
Subj	Iobj	669	794	3003	**0.81**
Iobj	Subj	86	153	806	0.19
Subj	Obl	318	31266	1106901	**0.96**
Obl	Subj	7409	10547	24001	0.04
Subj	Advmod	14559	15018	45122	**0.69**
Advmod	Subj	8684	15572	9331	0.31
Obj	Iobj	68	131	1020	0.19
Iobj	Obj	13	110	5062	**0.81**
Obj	Obl	5	3717	42274	0.46
Obl	Obj	814	9739	44494	**0.54**
Obj	Advmod	556	2446	27759	0.40
Advmod	Obj	1250	11140	33617	**0.60**
Iobj	Obl	45	48	1260	**0.52**
Obl	Iobj	105	266	878	0.48
Iobj	Advmod	44	39	461	0.31
Advmod	Iobj	187	327	705	**0.69**
Obl	Advmod	1664	7690	30813	0.39
Advmod	Obl	31	16322	46047	**0.61**

Vectorial representations

- Syntactic Structures; frequencies

- Syntactic Structures as a series of features.

Vectorial Representations

A	B	1_LP	2_LP	2_Mod	M1,2	1_MS	2_MS	LowP	Distr.
Obj	Iobj	1	1	0	1	1	1	0	0.009793
Obj	Iobj	1	0	0	1	1	1	0	0.018865
Obj	Iobj	0	0	0	1	0	0	1	0.146889
Obj	Obl	1	1	1	1	1	1	0	0.000435
Obj	Obl	1	0	0	1	1	0	0	0.036786
Obj	Obl	0	0	0	1	0	0	1	0.418376
Obj	Advmod	1	1	1	0	1	0	0	0.007243
Obj	Advmod	1	1	1	0	1	0	0	0.031862
Obj	Advmod	0	0	0	0	0	0	1	0.361596
Iobj	Obl	1	1	1	1	1	1	0	0.017294
Iobj	Obl	1	0	1	0	1	0	0	0.018447
Iobj	Obl	0	0	1	0	0	0	1	0.484243
Iobj	Advmod	1	1	1	0	1	0	0	0.024957
Iobj	Advmod	1	0	0	0	1	0	0	0.022121
Iobj	Advmod	0	0	0	0	0	0	1	0.261486
Obl	Advmod	1	1	0	1	0	1	0	0.016224
Obl	Advmod	1	0	0	1	0	1	0	0.074975
Obl	Advmod	0	0	0	1	0	0	1	0.300418

A	B	1_LP	2_LP	2_Mod	M1,2	1_MS	2_MS	LowP	Distr.
Obj	Iobj	1	1	0	1	1	1	0	0.009793
Obj	Iobj	1	0	0	1	1	1	0	0.018865
Obj	Iobj	0	0	0	1	0	0	1	0.146889
Obj	Obl	1	1	1	0	1	0	0	0.000435
Obj	Obl	1	0	0	0	1	0	0	0.036786
Obj	Obl	0	0	0	0	0	0	1	0.418376
Obj	Advmod	1	1	1	0	1	0	0	0.007243
Obj	Advmod	1	1	1	0	1	0	0	0.031862
Obj	Advmod	0	0	0	0	0	0	1	0.361596
Iobj	Obl	1	1	1	0	1	0	0	0.017294
Iobj	Obl	1	0	0	0	1	0	0	0.018447
Iobj	Obl	0	0	0	0	0	0	1	0.484243
Iobj	Advmod	1	1	1	0	1	0	0	0.024957
Iobj	Advmod	1	0	0	0	1	0	0	0.022121
Iobj	Advmod	0	0	0	0	0	0	1	0.261486
Obl	Advmod	1	1	1	1	0	0	0	0.016224
Obl	Advmod	1	0	0	1	0	0	0	0.074975
Obl	Advmod	0	0	0	1	0	0	1	0.300418

Results – Linear Regression

Waikato Environment for Knowledge Analysis, WEKA v.3.8.2 (Hall et al. 2009).

MOD (Theory α)

Correlation coefficient (*r)*	0.857
Element A targeting the Left Periphery	-0.3614,
Matching between A and B	**-0.0626**
Match subject and element A	-0.0575
Element B targeting the Left Periphery	-*0.0549*
Element A/B targeting a Low periphery	-0.1288

ARG (Theory β)

Correlation coefficient (*r)*	0.844
Element A targeting the Left Periphery	-0.3727,
Match subject and element A	-0.0577
Element B targeting the Left Periphery	-*0.0532*
Element A/B targeting a Low periphery	-0.1497

Genres, German varieties

- *Test asymmetries between genres/registers* (the problem of smaller datasets!)

- Other German varieties:

  e.g., UD treebank in Swiss German (UZH; *genres: blog, fiction, news, non-fiction, wiki; 100 trees; 1544 tokens;* Aepli & Clematide 2018)

- Micro- and Macro- variation in Germanic.

Further improvements

These results aim to add a quantitative dimension to the qualitative descriptions provided in cartographic studies.

I. Increase data sets.

II. Compare with experimental results (as “control groups” in the spirit of Gulordava et al. 2018).

III. Explore further statistical methods.

Danke!

References and questions: samo@blcu.edu.cn

北京语言大学语言学系
DEPARTMENT OF LINGUISTICS

Selected References

Gulordava K. et al. (2018) Colorless Green Recurrent Networks Dream Hierarchically. In NAACL, 1195–1205

Hall, M. et al. (2009) The Weka data mining software: An update. SIGKDD Explorations Newsletter 11(1). 10–18.

Merlo, P. & Ouwayda, S. (2018). Movement and structure effects on Universal 20 word order frequencies: A quantitative study. Glossa, 3(1), p.84.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In ITPCL (pp. 3-16). Springer, Cham

Rizzi, L., & Cinque, G. (2016). Functional categories and syntactic theory. Annual Review of Linguistics, 2, 139-163.

Roberts, I. (2004). The C-system in Brythonic Celtic languages, V2, and the EPP. In The Structure of CP and IP, 297-328

Samo, G. (2019a) A criterial approach to the cartography of V2, John Benjamins

Samo, G. (2019b) Cartography and Locality in German: a quantitative study with Dependency structures, RGG, 5, 1-2

Samo, G. & Merlo P. (2019) Intervention effects in object relatives in English and Italian: a study in Quantitative Computational Syntax. In Proceedings of Quasy, 46–56.

Samo, G, Zhao Y., Gamhewage G., (2020) Syntactic Complexity of Learning Content in Italian for COVID-19 Frontline Responders: A Study on WHO's Emergency Learning Platform, Verbum, 2020, vol. 11

Zeman, D. et al. (2020) Universal Dependencies 2.6. ÚFAL. Charles University. http://hdl.handle.net/11234/1-3226

Zimmermann, R. (2014) "Dating hitherto undated Old English texts based on text-internal criteria." Ms., University of Geneva.