# A cross-linguistic perspective on the realization of information structure

**Geert-Jan M. Kruijff**
*Computational Linguistics*
*University of the Saarland*
*Saarbrücken, Germany*
gj@coli.uni-sb.de

## 1. General setting

- Information structure is an *inherent* aspect of a sentential meaning, indicating its relation to the discourse context.
- Languages may *realize* IS in different ways. This differentiation holds not only *across* languages – within a single language, different means may *interact* to realize IS.
- This leads to various questions: How can we *predict* the use (or availability) of particular means to realize IS, and how strong a theory do we need to *formalize* such predictions?

## 2. Today's lecture

The **goal** for today's lecture:

Today we address the question of how one could *predict* the use (or availability) of particular means to realize information structure. We thereby focus on the use of word order as *structural indication of informativity*, and consider its interaction with tune. We investigate the question of what word order variability *is* and how to characterize informativity as a *typological category*.

### Material

- Chapter 6, *The category of informativity*
- "Formulating a category of informativity"

### Overview

1. **Typology and word order**: What is typology? What has been done about word order in typology? Steele's characterization of rigid, mixed and free word order. What does the data look like? What are the resulting hierarchies?
2. **Typology and information structure**: Criticism of existing approaches: No prediction! How could we define a *category of informativity*? The Null Hypothesis. Canonical $Rheme_{is}$-Foreground Position. Hypothesis I. $Rheme_{is}$-foreground projection ("focus projection") and the change of $Rheme_{is}$. Informativity Markedness Principle. Hypothesis II.
3. **Modelling competence**: What kind of perspective on grammar? How do we create multilingual grammars that capture the (effect of) typological implications?

# 3. Typology and word order

- **What is typology?**
  Typology is the study of languages aiming at the discovery of universal patterns that structure the systems of languages.
- **What is typology of word order?**
  - Patterning of *dominant* (or canonical) word order, (Greenberg, 1966; Hawkins, 1983).
  - Variability has been studied in (Steele, 1978) and e.g. (Skalička and Sgall, 1994), but there are no detailed hypotheses predicting variability.
- **A tripartite division**: Steele's rigid, mixed and free word order languages characterization.
  - Important: Mixed and free differ substantially in interaction of word order and tune in realizing information structure, and there are differences in fixation of word order.

## 3.1. Typology of dominant word order

- **Seminal work**: Greenberg's "Some universals of grammar with particular reference to the order of meaningful elements" (Greenberg, 1966). Hawkins' *Word Order Universals* (Hawkins, 1983) builds forth on Greenberg's work, and Venneman's.
- **Goal: Establish word order patterns**.
  - Patterning of ordering of verbal heads and their arguments, and of nominal heads and their arguments.
  - **Greenberg and Venneman**: uniform linearization for all different kinds of head/dependent constructions distinguishable in a language.
  - **Hawkins**: *word order types* are specific co-occurrence patterns, where each pattern shares a common (shared) feature functioning as typological indicator. With Hawkins, not just XV or VX but the entire pattern defines the word order type.
  - Implicational, statistical universals forming hierarchies.

## 3.2. Example hierarchies

- Hawkins presents a set of basic word order universals that extends Greenberg's original classification. From these universals, Hawkins derives two hierarchies for noun modifiers.

(1) Universal XIV, *Prepositional Noun Modifier Hierarchy*
Prep $\supset$ ((NDem $\vee$ NNum $\supset$ NA) & (NA $\supset$ NG) & (NG $\supset$ NRel))

(2) Universal XVIII, *Postpositional Noun Modifier Hierarchy*
Postp $\supset$ ((AN $\vee$ RelN $\supset$ DemN & NumN) & (DemN $\vee$ NumN $\supset$ GN))

- These hierarchies can be combined with universals on the relation between noun-modifier word order and the basic word order of the verb and its complements (primarily, subject and object) to explain the patterns leading to the 24 language types proposed by Greenberg (and further explored by Hawkins and colleagues).

## 3.3. Variability

- **Problem: There is no typology of variability**.
  - Skalička predicts "free" word order in the presence of morphologically rich languages, cf. (Skalička and Sgall, 1994; Sgall, 1995a). However, word order "freedom" is not defined.
  - Steele defines rigid, mixed and free word order but only suggests an indication for freedom depending on subject/verb agreement, (Steele, 1978). (Unfortunately, this fails on Turkish data about relative clauses, which need not have to have S-V agreement, (Hoffman, 1995).)
- **We want a typology of variability**
  - Forms that word order variability might take across languages.
  - Predictability of how information structure may be realized across languages.

### 3.4. Steele's definitions

- The characterization of a language's word order (matrix clauses, dependent clauses) as *rigid*, *mixed* or *free*. To determine variation, we use Steele's proposal: Of the constraints given below, if a language breaks constraint A (and hence A'), and B, then its word order is free, whereas its word order is rigid if none are broken. If some but not all constraints are broken, word order is mixed.

A. A variation on the basic word order in which the verb occurs in other than its position in the basic word order is to be avoided. (We understand "to be avoided" to mean that if the order would be grammatically *possible*, it would be highly marked.)

A'. A variation on the basic word order in which the verb occurs either initial or final to the clause is to be avoided, if the verb was neither initial nor final respectively in the basic order.

B. A variation on the basic word order in which the object precedes the verb and the subject follows the verb is to be avoided.

### 3.5. Example characterizations

- **Rigid**:

  X/XI: English
  IX: French, Italian

- **Mixed**: (B can be broken)

  XI: Dutch, German
  XXIII: Japanese

- **Free**: (both A and B can be broken)

  X: Czech, Russian
  XXIII: Hindi, Turkish

### 3.6. Typology of variability

- **Goal**: Establish implications that predict a degree of word order variability (rigid/mixed/free) on the basis of several basic aspects of a language's system.

- **Word order variability Null Hypothesis**:
  The more a language allows for the following phenomena to occur grammatically, the higher the likelihood that the language has a relatively free word order:

  i. frequent use of null anaphora,
  ii. lack of expletive pronouns,
  iii. a rich case system,
  iv. complex verbal morphology.

  Cf. work by Hale and by Speas referred to in (Kroeger, 1993), Skalička's work on language types, and Sgall *et al*'s comments in (Sgall et al., 1986).

- **More precisely**, we looked at 20+ languages (native informants mostly), investigating the following aspects:

  1. Greenberg/Hawkins word order type (Greenberg, 1966; Hawkins, 1983)
  2. Case strategies (Croft, 1990)
  3. Presence of articles, expletives, and pro-drop (Hale, Speas, Sgall *et al*)
  4. Verbal morphology (Skalička and Sgall, 1994)
  5. Subject-Verb agreement (Steele, 1978)

- **Hierarchies for VX, XV and SVO-based word order types**

### 3.7. The OV word order hypotheses

(3) Variation hypotheses for OV word order

a. OV-1:
OV & Strat(SubjObj,Case) & ((Agr(ErgActor,Verb) & Drop) ∨ (Agr(AbsSubj,Verb) & (VerbForm(Tense & (Aspect ∨ Mood ∨ Voice))))) ⊃ free

b. OV-2:
OV & Strat(SubjObj,Case) & (¬Agr(Subj,Verb) ∨ ¬Drop) ⊃ mixed

c. OV-3:
OV & ¬Strat(SubjObj,Case) & ¬Drop & ¬Agr(Subj,Verb) ⊃ rigid

- OV languages like Hindi, Hungarian, Korean and Turkish have a fairly rich verbal and nominal morphology. Skalička's typology correctly predicts free word order.
- Japanese on the other hand does have a nominal case system, but verbs only inflect for tense and there is no subject/verb agreement (except perhaps for honorifics). We understand mixed word order to arise from this combination of the presence of nominal case, and absence of rich verbal case – an observation confirmed by German dependent clause word order. (Note, German dependent clause word order rigidifies if nominal morphology is ambiguous.)
- Finally, the hypothesis correctly predicts the rigid word order of Dutch dependent clauses. The case marking is significant – prepositions realizing case do not make this word order mixed:

(4) a. ... de man wiens foto Kathy aan Elijah gaf.
... the man whose photo Kathy to Elijah gave.
"The man whose photo Kathy gave to Elijah."

b. * de man wiens foto aan Elijah Marie gaf.

### 3.8. Further OV hypotheses

(5) Further hypotheses on the basis of OV data

a. Articles:
OV & free ⊃ ¬(Art(def) ∧ Art(indef)), or
OV & (Art(def) & Art(indef)) ⊃ ¬free

b. Use of case:
Strat(NounModif,Case) ⊃ Strat(SubjObj,Case)

c. Expletives:
Agr(AbsSubj,Verb) & (VerbForm(Tense & (Aspect ∨ Mood ∨ Voice))) ⊃ ¬Expl

d. Pro-drop:
Agr(AbsSubj,Verb) & (VerbForm(Tense & (Aspect ∨ Mood ∨ Voice))) ⊃ Drop

### 3.9. VO word order hypotheses

For VO I lack sufficient data to come to a proper characterization. To cover the data, I propose the hypotheses below.

(6) Variation hypotheses for VO word order

a. VO-1:
VO & Strat(SubjObj,Case) & VerbForm(Aspect & Mood) ⊃ free

b. VO-2:
VO & ¬Strat(SubjObj,Case) & VerbForm(Tense & ¬(Aspect &Mood)) ⊃ rigid

### 3.10. SVO word order hypotheses

(7) Variation hypotheses for SVO word order

a. SVO-1:
SVO & Strat(SubjObj,Case) & (VerbForm(Tense & (Aspect ∨ Mood))) ⊃ free

b. SVO-1':
VerbForm(Tense & Aspect & Voice & Mood) ⊃ free

c. SVO-2:
SVO & Strat(SubjObj,Case) & (VerbForm(Tense & ¬(Aspect ∨ Mood))) ⊃ mixed

d. SVO-3:
SVO & ¬Strat(SubjObj,Case) & (VerbForm(Tense & ¬Mood)) ⊃ rigid

### 3.11. However ...

- SVO-2 does not predict mixed word order in Dutch matrix clauses!

(8) Revised variation hypothesis SVO-2:
SVO & (Strat(SubjObj,Case) ∨ (NG & ¬Strat(NomMod,L+Ad)) & (VerbForm(Tense & ¬(Aspect ∨ Mood)))) ⊃ mixed

- **Rigid**: English, French (no nominal case, even if rich verbal morph.!)
- **Mixed**: Dutch, German (primarily if nominal case system)
- **Free**: Czech, Russian

### 3.12. Further SVO hypotheses

(9) Further hypotheses on the basis of SVO data

a. Articles:
SVO & free ⊃ ¬(Art(def) ∧ Art(indef)), or
SVO & (Art(def) & Art(indef)) ⊃ ¬free,
SVO & (rigid ∨ mixed) ⊃ (Art(def) & Art(indef))

b. Use of case:
Strat(NounModif,Case) ⊃ Strat(SubjObj,Case)

c. Pro-drop:
free ⊃ Drop, rigid ∨ mixed ⊃ ¬Drop

### 3.13. Summary

- We presented Steele's characterization of (in)variability in word order as either rigid, mixed, or free.
- Based on empirical data from typologically different languages, I formulated several hypotheses that *predict* when a language has rigid, mixed, or free word order (thus improving on (Steele, 1978)).
- The empirical findings underlying the hypotheses confirmed to an important degree the predictions made by Skalička's (morphological) typology of languages (Skalička and Sgall, 1994; Sgall, 1995b), in showing the interrelation between morphology and variability.
- The proposal advanced here distinguishes itself from Skalička in discerning mixed word order besides rigid and free word order, and covering its predictability. This finer distinction is not vacuous. Mixed word order languages may realize information structure differently than either rigid or free languages. Unlike rigid languages, mixed languages can use word order to realize information structure, but not to the degree that free languages can. One consequence is that mixed languages have a more complex interaction between tune and word order than either rigid or free languages. On the other hand, whereas free word order languages usually lack articles, mixed do have them and are thus able to indicate contextual boundness that way.

# 4. The category of informativity

- **Predict** how a language (basically) is going to realize information structure.
- The approach we take:
  1. Present a Null Hypothesis, predicting the *canonical $Rheme_{is}$-Foreground position* ("canonical focus position")
  2. Consider degrees in changes of $Rheme_{is}$ – the *Informativity Markedness Hierarchy*
  3. Present hypotheses that how "non-canonical" $Rhemes_{is}$ can be realized

## 4.1. The informativity null hypothesis

(10) INFORMATIVITY HYPOTHESIS I (INITIAL VERSION)
In the unmarked case (unmarked mixed, free word order or unmarked tune), languages tend to realize (verbal) $Theme_{is}$ dependents/heads before $Rheme_{is}$ ones, backgrounded items before foregrounded items, and foregrounded dependents in canonical/systemic ordering.

- The formulation of the null hypothesis is nothing new. In different guises it appears throughout work in the Prague School, notably in FGD's Topic/Focus-articulation (Sgall et al., 1986; Hajičová et al., 1995; Hajičová, 1993) and its use of Firbas's communicative dynamism (Firbas, 1992).
- Despite its simplicity, the null hypothesis holds across a surprisingly large range of language types.

## 4.2. Examples of Informativity Hypothesis I

- For example, it holds for most of the OV languages we have considered so far. In rigidly verb-final languages like Japanese or Tamil, the $Rheme_{is}$-Foreground proper occupies the immediately preverbal position. In non-rigid verb-final languages like Sinhala the $Rheme_{is}$-Foreground proper can also be the postverbal dependent (Herring and Paolillo, 1995).
- The hypothesis holds also for German, which has a mixed OV word order at the dependent clause level, where the preverbal position (right before the verbal cluster) is usually considered to be the prefered position for the $Rheme_{is}$-Foreground (proper).

(11) *Japanese*

Taroo-wa [ susi-o ]$_F$ tabeta.
Taro-TOPIC susi-ACC eat-PAST

"Taroo ate the SUSHI."

(12) *Sinhala*

oya kaelaeae-we hiṭiyaa [ nariy-ek ]$_F$.
that forest-LOC live-PAST jackal-INDEF

"In that forest lived A JACKAL."
(Herring and Paolillo, 1995)(p.170)

- Similarly, we find that the SVO languages we have considered all tend to prefer to place the $Rheme_{is}$ towards the end of the clause (in the unmarked case). This holds particularly for the SVO languages with mixed word order, like Dutch or German, or free word order, like Czech or Russian.

### 4.3. Problematic case: Free OV languages

- Unfortunately, OV languages like Hindi, Hungarian or Turkish (XXIII) do not seem to obey Informativity Hypothesis I.
- The canonical *Rheme*$_{is}$ position in these languages is directly before the verb – wherever the verb is placed in the sentence.

(13) *Hungarian*

Mari JÁNOST látta.
Mary John-Acc see-Past

"Mary saw JOHN"

(14) *Turkish*

a. Esra kitab-ı okuyor
Esra book-ACC read-PRESPROG
"Esra is reading the BOOK."

b. Kitab-ı Esra okuyor
book-ACC Esra read-PRESPROG
"As for the book, it is Esra who is reading it."

- However, to (15) the hypothesis no longer applies, even though the word order as such is unmarked.

(15) *Hungarian*

Mari látta Jánost.
Mary see-Past John-Acc

"MARY saw John"

### 4.4. Revising the null hypothesis

- To revise the null hypothesis, we include four aspects:
  1. the dominant word order (XV/VX/SVO), as per the Greenberg/Hawkins typology;
  2. the degree of variability (rigid/mixed/free), as implied by the hypotheses presented earlier in this chapter;
  3. the presence or absence of a productive prosodic system (tune) and its interaction with the degree of variability; and,
  4. Venneman's idea of category consistency (cf. (Venneman, 1977; Hawkins, 1983)).

### 4.5. Category consistency

- Venneman's idea of category consistency is embodied in his "Natural Serialization Principle". This principle states that languages linearize all their operator-operand pairs *consistently*, thus either (strictly) operator before operand, or operand after operator. Particularly, it does seem that languages tend to have a canonical *Rheme*$_{is}$ position relative to the verbal head that is consistent with its dominant word order. This leads to (16).

(16) a. OV $\supset$ immediate preverbal position
b. VO $\supset$ postverbal position

### 4.6. Revised Informativity Hypothesis One

(17) Information Structure Hypothesis 1 (Canonical $Rheme_{is}$-Foreground Position):
The canonical $Rheme_{is}$-Foreground position (CRFP) is determined by either of the following hierarchies, depending on the language's type:

a. **OV/SVO-UFP-hierarchy**:
immediate preverbal position $\Subset$
(rigid $\vee$ mixed $\vee$ free) & SVO & (V2 $\supset$ VFinal)
$\vee$
(rigid $\vee$ mixed $\vee$ free) & $\neg$ tune & OV

b. **VO/SVO-UFP-hierarchy**:
postverbal position $\Subset$
(rigid $\vee$ mixed $\vee$ free) & SVO & $\neg$(V2$\supset$ VFinal)
$\vee$
(rigid $\vee$ mixed $\vee$ free) & $\neg$ tune & VO

### 4.7. Informativity Markedness Principle

- Sometimes, we may want to change the $Rheme_{is}$-Foreground proper – to avoid unwanted $Rheme_{is}$-Foreground projection, due to syntactic (or possibly even thematic) structuring, etc.
- For example, in an OV language the $Rheme_{is}$ could be preverbal but not *immediately* preverbal, in which case consistency would be maintained. However, if the $Rheme_{is}$ would not be immediately preverbal *nor* preverbal at all, then both Hypothesis I (describing the unmarked or canonical case) and consistency would be violated.

### 4.8. Changing the realization of $Rheme_{is}$

- From a viewpoint of economy, we could set up the following Informativity Markedness Principle. We use CC for category consistency, and FP for $Rheme_{is}$ position.

(18) Informativity Markedness Principle:
{ CC & UFP }
$\leq_{IM}$ { CC & non-canonical FP }
$\leq_{IM}$ { non-CC & non-canonical FP}

- Intuitively, if we would follow out economy, then we would also get the prediction that a language would first tend to the use its predominant strategy for realizing information structure, to obtain a more marked $Rheme_{is}$ position CC & non-canonical FP, (unless the construction would be ambiguous between a $Rheme_{is}$ proper realized in CFP and a marked FP). To obtain a really marked $Rheme_{is}$ position non-CC & non-canonical FP the language would resort to a different strategy, possibly in combination with its predominant strategy. Note that there may be different strategies for mixed and free word order languages. For example, mixed languages have articles at their disposal to realize contextual boundness, whereas free (in general) do not.

### 4.9. Informativity Structure Hypothesis 2

(19) Information Structure Hypothesis 2 (marked realization)

a. (CC & $\neg$UFP & $\neg$ambigous)
$\supset$ ((mixed $\vee$ free) $\supset$ word order) & (rigid $\supset$ tune)

b. ($\neg$CC & $\neg$UFP & (ambiguous $\vee$ $\neg$ambigous))
$\supset$ ((rigid $\vee$ mixed $\vee$ free) $\supset$ word order & tune)

- Naturally, a more marked realization must be grammatical.

### 4.10. Structural indications of informativity in OV

- Examples:
  1. Rigidly and non-rigidly verb-first OV languages have an immediately post-verbal unmarked $Rheme_{is}$ position.
  2. Rigid OV realize IS using predominantly tune; mixed and free OV languages use predominantly word order.
  3. Non-rigidly verb-initial OV languages can have a marked immediately preverbal $Rheme_{is}$.
  4. Rigidly and non-rigidly verb-initial OV languages with mixed or free word order can have marked $Rheme_{is}$ position towards the end of the sentence, using just word order unless the structure as such would be ambiguous between a CRFP construction. To avoid ambiguity, tune is then used.

## 5. Modelling competence

- The bare necessities:
  1. Linearization up to free word order
  2. Syntactic dependency, cf. (Greenberg, 1966; Venneman, 1977; Hawkins, 1983)
  3. Linearization controlled by (a semantic notion of) information structure, syntax/semantics-interface.
- Generative power up to mild-context sensitivity, possibly higher; cf. (Hoffman, 1995). Certainly stronger than CFG: Discontinuous/long-distance dependencies cannot be covered in CFGs. (A 'classical' CF-PSG can only cover continuous word order variation through enumeration.)
- Many different proposals: TAG/LTAG, categorial grammar (CTL, CCG/MCCG/Set-CCG, ...), HPSG (domains – Reape, Kathol, Meurers, ...), etc.
- We focus on categorial grammar, because of its transparent syntax/semantics-interface (and well-defined logical system).

### References


William Croft. 1990. *Typology and Universals.* Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, United Kingdom.

Jan Firbas. 1992. *Functional Sentence Perspective in Written and Spoken Communication.* Cambridge University Press, Cambridge, United Kingdom.

Joseph H. Greenberg. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Language*, pages 73–114. The MIT Press, Cambridge, Massachusetts, second edition edition.

Eva Hajičová, Hana Skoumalová, and Petr Sgall. 1995. An automatic procedure for topic-focus identification. *Computational Linguistics*, 21(1):81–94, March.

Eva Hajičová. 1993. *Issues of sentence structure and discourse patterns*, volume 2 of *Theoretical and Computational Linguistics.* Charles University Press, Prague, Czech Republic.

John A. Hawkins. 1983. *Word Order Universals.* Academic Press, New York, London, etc.

Susan C. Herring and John C. Paolillo. 1995. Focus position in SOV languages. In Pamela Downing and Michael Noonan, editors, *Word Order in Discourse*, volume 30 of *Typological Studies in Language*, pages 163–198. John Benjamins, Amsterdam, The Netherlands.

Beryl Hoffman. 1995. *Computational Analysis of the Syntax and Interpretation of "Free" Word-Order in Turkish.* Ph.d. thesis, ircs report 95-17, University of Pennsylvania.

Paul Kroeger. 1993. *Phrase Structure and Grammatical Relations in Tagalog.* Dissertations in Linguistics. CSLI Press, Stanford, CA.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* D. Reidel Publishing Company, Dordrecht, Boston, London.

Petr Sgall. 1995a. Prague school. In Jan-Ola Östman Jeff Verschueren and Jan Blommaert, editors, *Handbook of Pragmatics (Manual)*, pages 429–435. John Benjamins, Amsterdam, The Netherlands.

Petr Sgall. 1995b. Prague school typology. In Masayoshi Shibatani and Theodora Bynon, editors, *Approaches to Language Typology*, pages 49–84. Clarendon Press, Oxford, United Kingdom.

Vladimír Skalička and Petr Sgall. 1994. Praguian typology of languages. In Philip A. Luelsdorff, editor, *Prague School of Structural and Functional Linguistics*, number 41 in Linguistics and Literary Studies in Eastern Europe, pages 333–357. John Benjamins, Amsterdam, The Netherlands.

Susan Steele. 1978. Word order variation: A typological study. In Joseph H. Greenberg, editor, *Universals of Language. Volume 4: Syntax*, pages 585–624. Stanford University Press, Stanford, California.

Theo Venneman. 1977. Konstituenz und Dependenz in einigen neueren Grammatiktheorien. *Sprachwissenschaft*, 2:259–301.