



Subprojekt A3: *Suboptimal Syntactic Structures*

Project Leader: *Wolfgang Sternefeld*

Concerns itself with:

- Marginal and ungrammatical structures
- Data types and syntax theory
- Judgements as a data type
- Takes a generative perspective

<http://www.sfb441.uni-tuebingen.de/a3/>

Judgements in syntax:

Why they are good, how they can be better

Sam Featherston

SFB441: Linguistic Data Structures

Tübingen University

DGfS, Mainz, 25.2.2004

3

Talk outline

- 1 Judgements in syntax: the challenges
- 2 Improving judgement data quality
 - Developing experimental methods
 - Distinguishing syntax-relevant effects
- 3 Psycholinguistics of judgements
 - Categorical judgements vs relative judgements
 - Judgement data vs frequency data
- 4 A standard scale for judgements
- 5 Project A3's database of judgements

Linguists' judgements: the data type

4

Judgements are fuzzy (Schütze 1996)

- variable
- not replicable

Judgements are difficult to communicate

- no standard scale

Uncertainty about the construct measured

- contributing factors
 - * context
 - * meaning
 - * form
- Tested object: actually a form-meaning pair

Linguists' judgements: their use

Theories of syntax abstract from the data

- Abstraction distorts the data
- No standard abstraction

Conflicting conceptions of what judgement data is evidence of

- Good vs bad
- Good vs marginal_i ... marginal_n vs bad
- "better than"

We need judgement data too

- Maintain contact with previous work
- Its phenomena require explanation
- Close correlation to other data types
- It can be hardened as a data type
- It actually has advantages over other types

So should we abandon judgements?

Corpus data has many advantages

- Accessible
- Developed technology
- Hard data
- "Real" language

Processing data has advantages

- Reflects mind-internal effects
- Developed technology
- Hard data
- Permits strict control and focus

The challenges for judgements

1: Judgements are fuzzy

Hardening judgements as a data type

2: Multiple contributory factors

Distinguishing grammar and other factors

3: Poor understanding of judgements

Distinguishing types of judgements

4: Contrast with frequency data

Distinguishing judging and production

5: Communicating judgements

Making judgements inter-subjective

Hardening judgements as a data type

Multiple informants

- evens out variation
- removes bias
- greater accuracy

Multiple lexicalizations of each structure

- controls for irrelevant effects
- allows claim effects are structure-related
- can sharpen results

Design

- control conditions
- comparison set

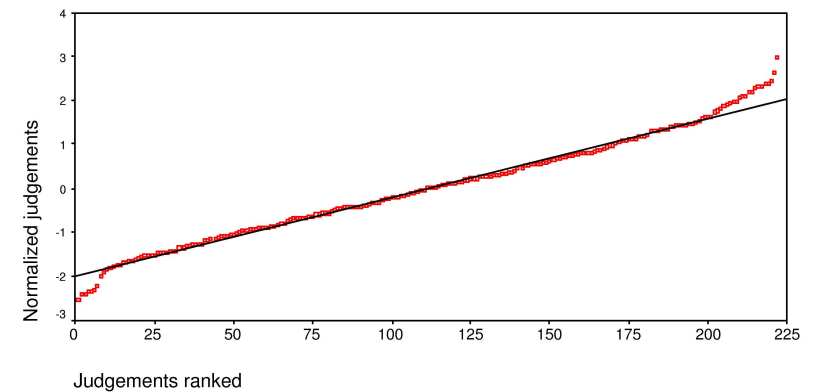
Experimental harder judgements

Magnitude estimation (Bard et al 1996)

Elicited grammaticality judgements, but ...

- judgements are anchored by a reference item
- subsequent judgements are relative to previous ones
- judgements are proportional
- the scale is open-ended
- the scale has no minimum division

The pattern of MagEst results



Keller's (2003) power law for judgements not supported

Poulson (1986): we can judge difference not ratio

Thermometer scale judgements

Relative judgements on a linear scale give the same results as MagEst.

- Exclude magnitude from instructions. Instead:
"If this is a 20 and that is a 30, what do you give this one?"
- 2 reference examples (or comparison set)
(First fixed point sets anchor, second one sets amplitude)
- Thermometer scale
(= open-ended, no minimum division, two fixed points)

Distinguishing syntax-relevant effects

Abstracting from the data

We like:

Grammaticalness vs Acceptability

- "... ideal speaker-listener, ..." etc Chomsky (1965)
- a psycholinguistic criterion

We don't like:

Criterion of categorical ungrammaticality

- "grammaticality" vs "markedness"
- a product of abstraction to binary grammaticality

Re-thinking syntax relevance

Rule 1 An effect is not syntax-relevant if it can be accounted for by known performance or processing factors.

- * Heaviness
- * Information structure
- * Frequency
- * Processing complexity
- * Phonotactics
- * Plausibility

Rule 2 Any effect is syntax-relevant if it cannot be thus accounted for.

Implications for data

Identifying syntax-relevant information is like peeling an onion blindfolded. You have to strip away layer after layer but you never know when you can stop.

So....data types with least irrelevant processing content are best for syntax.

On-line processing data has proportionally least syntax-relevant information.

Judgements and frequency have proportionally more syntax-relevant information.

Two types of judgements?

Relative judgements:

- Continuum of well-formedness
- No clear upper or lower bound

Categorical judgements:

- Intuition of categorical grammaticality
- "It's better than the other but I would never say it"

Distinguishing types of judgements

Categorical judgements and relative judgements

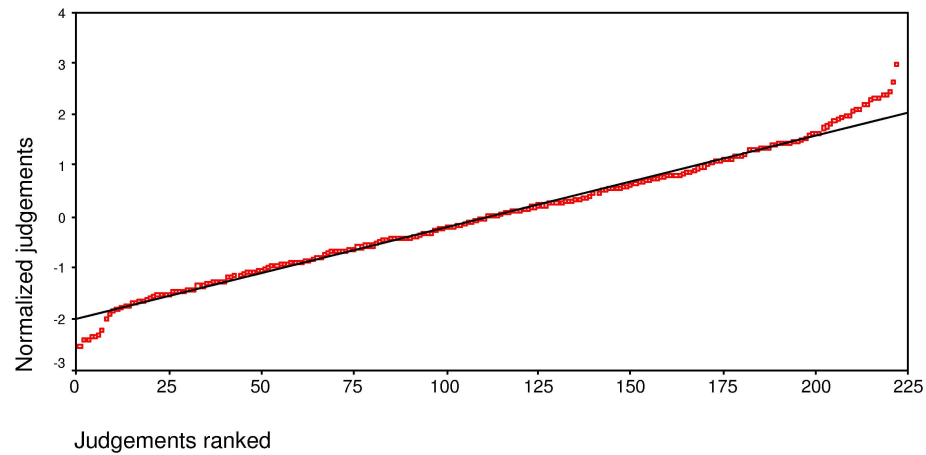
Categorical judgements

- Reflect occurrence
- Closer to corpus data
- Scale with fixed end points
- Triggered by:
 - * Single judgements
 - * Unrelated structures

Relative judgements

- Reflect the formants of well-formedness
- Continuum
- Requires a comparison set for transferability
- Triggered by:
 - * Multiple linked judgements
 - * Reference to own previous judgements

Relative judgements and frequency data



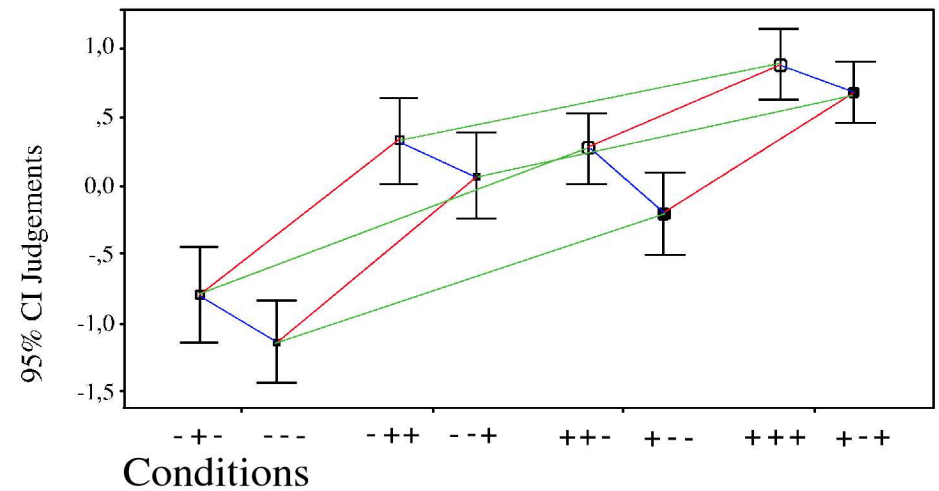
23

The contrasting data patterns: judgements

Relative judgements reveal (Keller 2000):

- a continuum of well-formedness
- constraints with specific, quantifiable violation costs
 - * cumulative
 - * applied blindly and exceptionlessly
 - * and are *survivable*

Cumulative and quantifiable violation costs



24

The contrasting data patterns: frequency

Frequency data reveals:

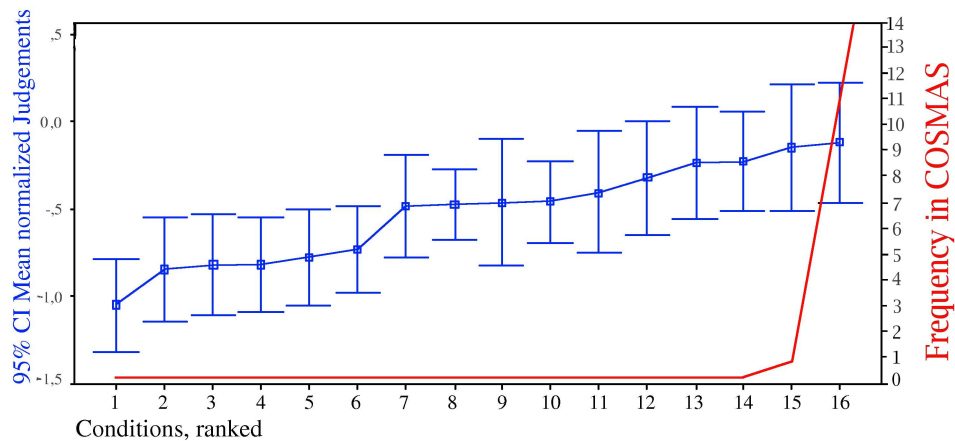
- competitive interaction
- probabilistic functioning
- dichotomous output: occurrence vs non-occurrence

Corpus data exhibits probabilistic competition

COSMAS, IDS, Mannheim (531 million word forms)

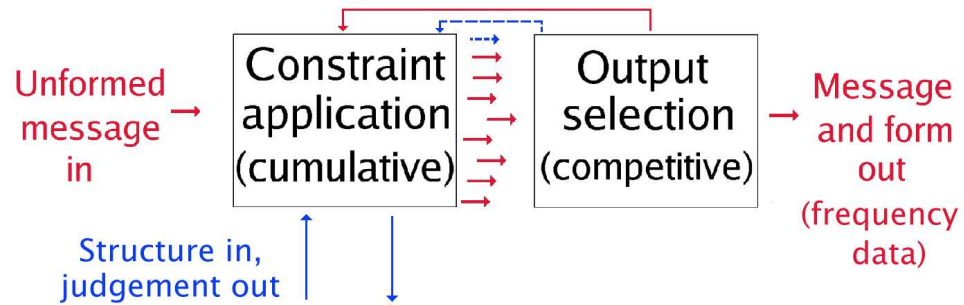
ihn _i ihm _i	("him.ACC him.DAT")	0 hits
ihn _i ihm _i selbst	("him.ACC him.DAT SELF")	0 hits
ihm _i ihn _i	("him.DAT him.ACC")	0 hits
ihm _i ihn _i selbst	("him.DAT him.ACC SELF")	0 hits
ihm _i sich _i	("him.DAT REFL.ACC")	0 hits
ihm _i sich _i selbst	("him.DAT REFL.ACC SELF")	0 hits
ihn _i sich _i	("him.ACC REFL.DAT ")	1 hit
ihn _i sich _i selbst	("him.ACC REFL.DAT SELF")	14 hits

The contrast of frequency and judgement patterns 27



The Decathlon Model

The Decathlon Model



Two modules in syntactic processing

Constraint Application

- cumulative
- blind and exceptionless
- assigns violation costs
- outputs well-formedness weightings

Output Selection

- competitive
- probabilistic
- excludes candidates

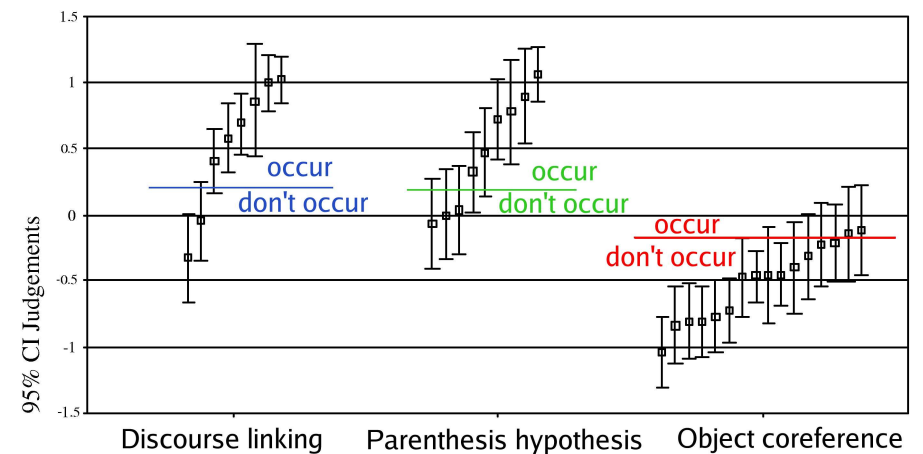
The contrast of relative judgements and frequency-based data:

- Frequency data reflects both modules
- Judging is neither input nor output processing
- Relative judgements are output of Constraint Application

Evidence for the Decathlon Model

- A single module predicts that well-formedness determines output.
- Decathlon Model claims that competition determines output, so....
- No single level of well-formedness triggers occurrence in the output.

Output is driven by competition



Answer: in the bag of constraints in Constraint Application

This does not exclude a separate innate(?) grammar.

Judging is assessing computational complexity.

Symptom of learned constraints (eg frequency, lexical preference):

> computational complexity

Symptom of perceptual, conceptual, architecture-based constraints:

> computational complexity

Lack of external difference doesn't exclude different origin.

Implication for syntax: grammar resembles judgements 35

- 1 Constraints have quantifiable and cumulative a violation costs.
- 2 Constraints are applied exceptionlessly (ie NOT OT violability).
- 3 But constraints are *survivable* (ie constraint application is separate from output selection).
- 4 Well-formedness is a continuum, but absolute, not relative.
- 5 Competition is in output selection, not in interaction of constraints.

So... Optimality Theory is a model of production, not of the grammar. Neither relative well-formedness nor competition are features of the narrow syntax. Every constraint always affects well-formedness.

All empirically adequate grammars need to be equipped with an extra module which specifies violation costs. So....

Pinker (1999): Words and rules

Featherston: Words, rules and sanctions

- 1 Categorical judgements and frequency reflect output.
- 2 Relative judgements best reveal the formants of well-formedness.
- 3 For syntax, relative judgements are the richest information source.

Syntax and data: a way forward

Theory:

Current syntax doesn't match the data.
Gradiance massively complicates theory.

Data:

Judgements must be possible without experiments.

Communicating judgements:

Judgements are not inter-subjective.
Experimental judgement data is non-transferable

Solution:

A scale of well-formedness with 5(?) cardinal values.
(like "cardinal vowels")

Advantages of a standard scale of well-formedness

With a scale anchored by standard examples...

- Individual judgements become inter-subjective (like cardinal vowels for phoneticians).
- Finer judgements become possible without experiments (cf Celsius, km per hour, known vs unknown currency)
- Experiments can use them as fillers and become transferable
- An abstracted, manageable, 5-step gradiance can enter theory. (A Trojan horse for the next ...)
- Gradiance brings violation costs, survivability, cumulativity etc

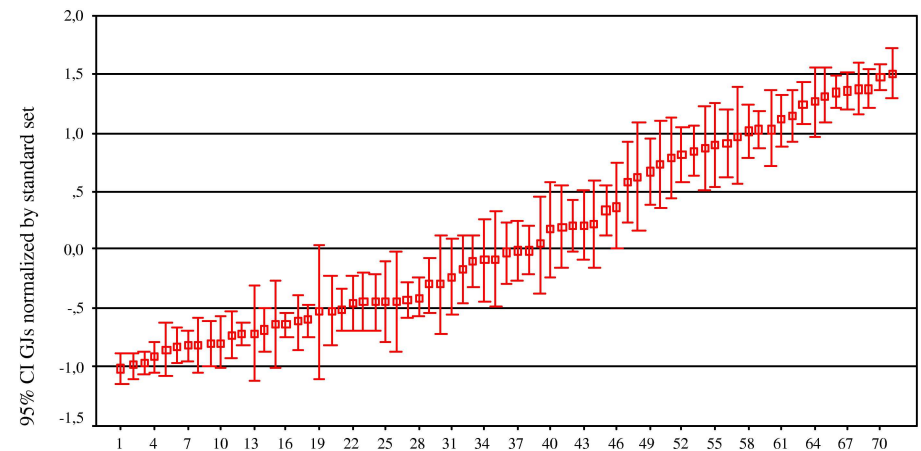
Creating a reference set

71 varied examples tested in two judgement experiments

Some sample rankings from the results:

- 1 Der Dirigent bedauert er, dass den Sänger überfordert hat.
- 2 Der Patient hat das geldgierige Zahnarzt überlistet.
- 6 Auf der Insel Helgoland regnet sie 200 Tage im Jahr.
- 16 Welches Zimmer weißt du nicht wo sich befindet?
- 21 Die Bergführer haben ihn einander als Begleiter empfohlen.
- 23 Der Stürmer hat dem Torwart vorsätzlich gefoult.
- 28 Der Komponist hat dem neuen Tenor es zugemutet.
- 42 Was ich wissen will, ist wen wer in dieser Affäre betrügt.
- 70 Nach der Arbeit gehe ich gern mit Freunden ins Kino.

Distribution of results



Can people distinguish five cardinal well-formedness values?

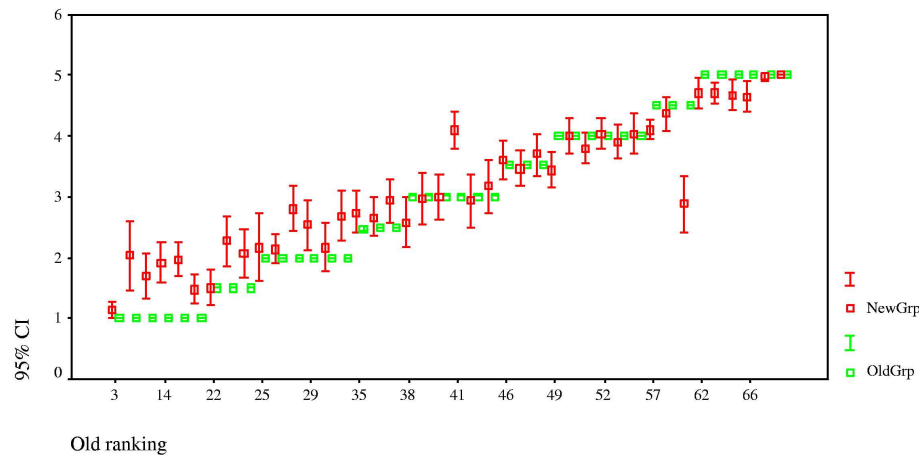
Procedure

1. Five groups presented to subjects - two examples each
2. Practice: Five practice sets of five items at a time, one of each group.
3. Experiment: Items presented one by one for group assignment

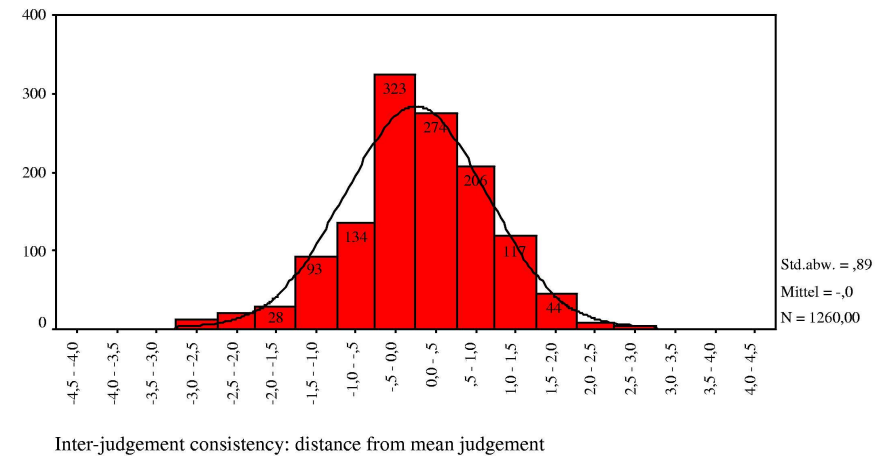
Two tests of success:

- Do judges agree with previous group assignments?
- Do judges agree with each other?

Matching new examples to groups



Intersubject consistency: Variance around means



- Gruppe A: In der Mensa essen viele Studenten zu Mittag.
- Gruppe B: Sie hofft, das Finanzamt hat den Betrüger überlistet.
- Gruppe C: Ich habe dem Kunden sich selbst im Spiegel gezeigt.
- Gruppe D: Der Komponist hat dem neuen Tenor es zugemutet.
- Gruppe E: Der Händler glaubt er, dass den Politiker bestochen hat.

- 1 Wen glaubt er, hat der Student mit seinen Fragen verärgert?
- 2 Wem hat der Förderer welche Subvention bewilligt?
- 3 Da gerechnet mit hat der Franz natürlich nicht.
- 4 Ich habe ihm ihn selbst im Spiegel gezeigt.
- 5 Der Agent hat ihm den ganzen Plan verraten.

Five group gradience results

45

Raw results, errors included

Assignments to previously obtained groups

- Direct hits 72.8%
- Direct hit or one off 89.7%

Judgements distance from all-subjects mean:

- Up to one group off 75.7%
- Up to 1.5 groups off 91%

We still need to do some work to find the most reliable cardinal group exemplars.

Summary

47

Judgements can be hard data if simple methods of control applied.

Thermometer scale judgements perhaps better than MagEst.

We can distinguish relative judgements and categorical judgements.

Relative judgements are most informative about the components of the grammar.

Judgements relate to frequencies as in the Decathlon Model

A standard scale provides an inter-subjective anchor for judgements.

A standard scale would permit gradient syntax without experiments

Syntax needs a deposit of judgements to implement peer review.

46

Project A3's database of independent judgements

Judgements must be in principle replicable

Central data base of judgements from the literature

With judgements from independent informants

Data base should be commentable

Published works should submit judgements

Pilot version: <http://www.sfb441.uni-tuebingen.de/a3/db/>

48

Thank you!