CITIZEN SCIENCE IN LINGUISTICS: PAST, PRESENT AND FUTURE

Sabine Stoll
University of Zurich
Department of Comparative Linguistics
Tracing the roots of syntax with Bayesian phylogenetics

Luke Maurits1 and Thomas L. Griffiths
Department of Psychology, University of California, Berkeley, CA 94704

Edited by Richard N. Aslin, University of Rochester, Rochester, NY, and approved July 25, 2014 (received for review October 8, 2013)

The ordering of subject, verb, and object is one of the fundamental components of the syntax of natural languages. The contribution of basic word order across the world’s languages is highly non-uniform with the majority of languages having either subject-object-verb (SOV) or subject-verb-object (SVO). Explaining this fact the explanatory adequacy of any putative theory of word-order functionality.

Some authors have suggested that all present languages are descended from a common ancestor with SOV word order (10, 11). This hypothesis is interesting in light of results suggesting

LETTER

do:10.1038/nature09923

Evolved structure of language shows lineage-specific trends in word–order universals

Michael Dunn1,2, Simon J. Greenhill3,4, Stephen C. Levinson1,2 and Russell D. Gray3

Evolutionary dynamics of language systems

Simon J. Greenhill5,6, Chieh-Hsi Wu7, Xia Hu8, Michael Dunn9, Stephen C. Levinson3,9, and Russell D. Gray5,8

Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement

R. D. Gray,1 A. J. Drummond,2 S. J. Greenhill1

www.sciencemag.org SCIENCE VOL 323 23 JANUARY 2009

A comparison of worldwide phonemic and geneti variation in human populations

Nicole Creanza2, Merritt Ruhlen2, Trevor J. Pemberton1, Noah A. Rosenberg9, Marcus W. Feldman4,7, and Sohini Ramachandran4,7,9

This contribution is part of the special series of Inaugural Articles by

Cortical tracking of hierarchically linguistic structures in connected speech

Nai Ding1,2, Lucia Melloni1,3,5, Hang Zhang1,3,5, Xing Tian1,9,10, & David Poeppel1,11

Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language

Dan Dediu1,2,6, Rick Janssen3, Scott R. Moisik1,2

Grammars are robustly transmitted even during the emergence of creole languages

Damian E. Blasi1,2, Susanne Maria Michaelis2,3 and Martin Haspelmath2,3

Sound–meaning association biases evidenced across thousands of languages

Damian E. Blasi1,2,4, Soren Wichmann5,6, Harald Hammarstrom7, Peter F. Stadler8,9, and Morten H. Christiansen10

The Neurophysiology of Language Processing Shapes the Evolution of Grammar: Evidence from Case Marking

Balthasar Bickel1,4, Alena Witzlack-Makarevich2, Kamal K. Choudhary3, Matthias Schlesewsky4,5, Ina Bornkessel-Schlesewsky6,6

Is “Huh?” a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items

Mark Dingemanse1,*, Francisco Torreira1, N. J. Enfield1,2

Kinship Categories Across Languages Reflect General Communicative Principles

Charles Kemp6,7 and Terry Regier2

www.sciencemag.org SCIENCE VOL 336 25 MAY 2012
HUGE UNRESOLVED QUESTION:
How can language be acquired by children?
EXTREME LINGUISTIC VARIATION
EXTREME CULTURAL VARIATION
EXTREME INDIVIDUAL VARIATION

All Data (n = 4867)

Size of Productive Vocabulary versus Age (months)

Quantile

Source: Wordbank
WE NEED:

1. Data on as many diverse languages and cultures as possible

2. Data on as many children as possible recorded in their natural context
SOLUTION: CITIZEN SCIENCE
BIG CITIZEN SCIENCE PROJECT
IN A SMALL COMMUNITY
IN THE FIELD:

LANGUAGE ACQUISITION IN CHINTANG
LINGUISTIC ANALYSIS

lexicon  orthography  grammar

development  analysis

with citizens
DATA PREPARATION

recordings  metadata collection  transcription

with and by citizens
ANNOTATIONS

grammatical annotations

situational annotations

gestural annotations

by linguists
Linguistic Analysis

- lexicon
- orthography
- grammar

with and by citizens

Data Preparation

- recordings
- metadata
- transcription

with and by citizens

Annotations

- grammatical
- situational
- gestural

by linguists
OUR DILEMMA

Recording
1 hour → 1%

Video linking
10-20 hours → 10 - 11%

Transcription & Translation
30-100 hours → 33 - 50%

Grammatical annotation
50-80 hours → 40 - 55%

1 hour of recording
90-200 hours of data preparation

Corpus size: 250 h of recording
THE CHINTANG TEAM AND FUNDING

PIs Balthasar Bickel & Sabine Stoll

Ban jade, Goma
Bibiko, Hans-Jörg
Bickel, Balthasar
Berthold, Falko
Cannedo, Laura
Domberg, Andreas
Erlach, Michael
Ghimire, Sandesh
Gaenszle, Martin
Hegetschei ler, Shirin
Harbodt, Silke
Kantak, Nicole
Klein, Felix
Krämer, Tatjana
Labs, Kristina
Lieven, Elena
Lorenz, Nicole
Paudyal, Netra
Pettigrew, Judith
Polka u, Claudia
Poppitz, Sindy
Prokopchuk, Joel
Rai, Chandra
Rai, Anita
Rai, Daya
Rai, Durga B.
Rai, Durga K.
Rai, Ganesh
Rai, Janaki
Rai, Lash K.
Rai, Rikhi
Rai, Novel K.
Rai, Shanti M.
Rai, Manoj
Rai, Rabindra
Sanjuan, Marcel
Sauppe, Sebastian
Schackow, Diana
Schikowski, Robert
Schneider Sarah
Seeg, Jenny
Strassberg, Joel
Stoll, Sabine
Visiennon, Kodjo
Washington, Kerstin
Weymuth, Rachel
Wienholz, Anna
Wohlgemuth, Jan
Wolters, Kristin
Wunder, Saskia
Zakharko, Taras
Zeiske, Mathias
Situation now

- Small sample
- Extreme data preparation efforts

Envisioned future situation

- Big sample
- Less data preparation efforts
involve more citizens

DATA PREPARATION

recordings metadata transcription

involve more citizens

ANNOTATIONS

grammatical situational gestural

involve citizens
BUT HOW TO GET REALLY BIG DATA ?
Amazon Mechanical Turk (MTurk) operates a marketplace for work that requires human intelligence. The MTurk web service enables companies to programmatically access this marketplace and a diverse, on-demand workforce. Developers can leverage this service to build human intelligence directly into their applications.
CITIZEN SCIENCE PROJECTS

Shakespeare’s World Project:

transcribe handwritten documents by Shakespeare’s contemporaries
CITIZEN SCIENCE PROJECTS

Project SAC-KOKOS:

an online platform to manually correct OCR errors
The Wolf and the Crane (Aesop)

A Wolf had been gorging on an animal he had killed, when suddenly a small bone in the meat stuck in his throat and he could not swallow it. He soon felt terrible pain in his throat, and ran up and down groaning and groaning and seeking for something to relieve the pain. He tried to induce every one he met to remove the bone. "I would give anything," said he, "if you would take it out." At last the Crane agreed to try, and told the Wolf to lie on his side and open his jaws as wide as he could. Then the Crane put its long neck down the Wolf’s throat, and with its beak loosened the bone, till at last it got it out.

Not mentioned before

This word is non-referring

Done
GAMIFICATION in the LANGUAGE SCIENCES
– a longstanding enterprise –
CITIZEN SCIENCE PROJECTS AND GAMIFICATION

help scientists to identify human proteins within human cells

Project Discovery
MY VISION

A digital laboratory to research child-caretaker interactions in a quasi-naturalistic environment

Collaboration with:
FUTURE PLANS: DIGITAL LAB

- Stimuli
- Data preparation
- Gamification
- Data analysis
- Feedback

- Speech production and recording
- Transcription
- Social networks
Thank you very much!