# What is complexity?

**Kaius Sinnemäki**
(sinnemaki (att) gmail.com)
University of Helsinki

*Formal linguistics and the measurement of grammatical complexity*

Seattle, WA, March 24, 2012

# Overview of the talk

1. General background
2. A cross-linguistic approach
3. A case study
4. Conclusion

- The slides are available on my website: www.ling.helsinki.fi/~ksinnema/

# 1. General background 1/4

- There has been a lot of research on complexity and complex systems in the natural sciences, economics, social sciences, and now also increasingly in linguistics.
- However, there is no consensus over the formulation of the notion of *complexity* in the science(s) of complexity.
  - What is going on here? Is this a symptom of a young discipline or a more fundamental issue? Probably the latter.
  - Mikulecky (2001: 344): Complexity is "the property of a real world system that is manifest in the inability of any one formalism being adequate to capture all its properties."
- Edmonds (1999) and Lloyd (2001) provide lists of different formalisms that have been used for measuring complexity, each about 40 entries long.

# Edmonds (1999)

Abstract computational complexity; **Algorithmic information complexity;** Arithmetic complexity; **Bennett's 'logical depth';** Cognitive complexity; **Connectivity;** Cyclomatic number; **Descriptive/interpretative complexity;** Dimension of attractor; **Ease of decomposition;** Economic complexity; **Entropy;** Goodman's complexity; **Horn complexity;** Information; **Information gain in hierarchically approximation and scaling;** Irreducibility; **Kemeny's complexity;** Length of proof; **Logical complexity/arithmetic hierarchy;** Loop complexity; **Low probability;** Minimum number of sub groups; **Minimum size;** Mutual information; **Network complexity;** Number of axioms; **Number of dimensions;** Number of inequivalent descriptions; **Number of internal relations;** Number of spanning trees; **Number of states in a finite automata;** Number of symbols; **Number of variables;** Organised/disorganised complexity; **Shannon information;** Simplicity; **Size;** Size of grammar; **Size of matrix;** Sober's minimum extra information; **Sophistication;** Stochastic complexity; **Syntactic depth;** Tabular complexity; **Thermodynamic depth;** Time and space computational complexity; **Variety.**

# Lloyd (2001)

Algorithmic information content; **Algorithmic mutual information;** Channel capacity; **Chernoff information;** Code length; **Computational complexity;** Conditional algorithmic information content; **Conditional information;** Correlation; **Cost;** Crypticity; **Dimension;** Effective complexity; **Effective measure complexity;** Entropy; **Excess entropy;** Fisher information; **Fractal dimension;** Grammatical complexity; **Hierarchical complexity;** Homogeneous complexity; **Ideal complexity;** Information; **Information-based complexity;** Lempel-Ziv complexity; **Logical depth;** Metric entropy; **Minimum description length;** Mutual information; **Organization;** Renyi entropy; **Schema length;** Sophistication; **Space computational complexity;** Stochastic complexity; **Stored information;** Thermodynamic depth; **Time computational complexity;** Topological epsilon-machine size; **Tree subgraph diversity;** True measure complexity.

# 1. General background 2/4

- However, these formalisms share so much in common that they could be informally classified in a few groups only (Lloyd 2001; Rescher 1998: 8–16):
  1) difficulty of creation (or generation),
  2) difficulty of description,
  3) degree of organization.

  →Possible to characterize complexity at a general level.

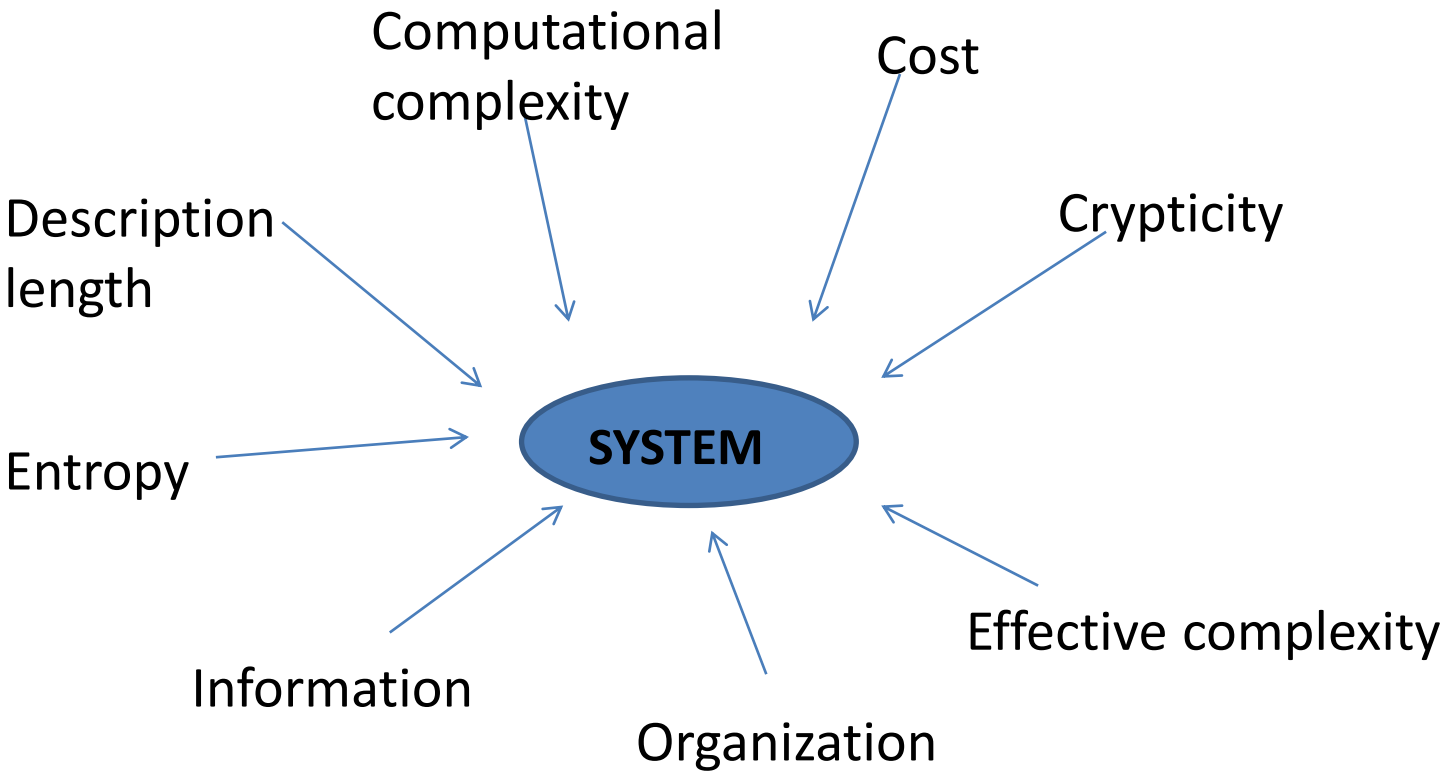
- Page (2011: 31–33) conflates these even further and describes complexity with just two general properties:
  1) it is not easily described, evolved, engineered, or predicted,
  2) it lies between order and randomness (or disorder).

# 1. General background 3/4

- At a general level, I characterize complexity as **the number and variety of parts and their interactions** (Simon 1996: 183–184; Rescher 1998: 1).
  - A system with many interacting parts is both highly organized and difficult to describe.

- Note how close these characterizations come to everyday language use. For instance, Oxford Advanced Learner's dictionary provides two senses for the adjective *complex*:
  1. consisting of many interrelated parts (= degree of organization)
  2. and being difficult to understand (= difficulty of description).

# 1. General background 4/4




- Different formalism provide different and partial windows to a system's complexity.

# Overview of the talk

1. General background
2. A cross-linguistic approach
3. A case study
4. Conclusion

# 2. A cross-linguistic approach

- I propose that we need a few general criteria in order to compare grammatical complexity across languages (Sinnemäki 2011):
  1. grammatical complexity is separated from efficiency (or difficulty / cost, cognitive complexity),
  2. local complexity is separated from global complexity,
  3. complexity is broken down into different types,
  4. complexity is formally measured as the description length of an object's structure (however that is realized).
  - linguists' tools provide a feasible starting point.
- These criteria constrain the study of grammatical complexity to a certain type of complexity of a certain local pattern → probably easier to study how grammatical complexity might correlate with difficulty.

# 2.1 Complexity vs. difficulty 1/2

- A metric of grammatical complexity should not be based on difficulty but kept apart from it (e.g., Dahl 2004). Why?
- If our metric of grammatical complexity was based on cognitive complexity (cf. Kusters 2003), three problems would arise for a general cross-linguistic approach.

1. Different user-types (speaker, hearer, first language acquirer, second language learner) may experience a linguistic pattern differently, which results in conflicting complexity measures.
   - For instance, redundant agreement may be useful to first language acquirers but costly to adult L2 learners (Kusters 2003).
   - *Finnish:* *piene-**ssä*** *punaise-**ssa*** *talo-**ssa***
     small-iness. red-iness. house-iness.
     'in a small red house'

# 2.1 Complexity vs. difficulty 2/2

2. A user-based approach would require focusing on one user-type over the others or defining an idealized user-type.

   →Loss of variation.

3. Complexity is just one factor affecting efficiency:

   →There are other ways to demonstrate how grammatical and cognitive complexity might correlate other than trying to equate the two from the outset (e.g., Hawkins 2004).

# 2.2 Local vs. global complexity

- Local complexity = the complexity of some part of a system (e.g., syllable, inflectional paradigms).
- Global complexity = the overall complexity of a system (e.g., the grammar of a language).
- Two main problems why measures of global complexity are unattainable (Miestamo 2008):
  1. Global complexity requires comprehensive descriptions; impossible due to the infinitude of grammar (cf. Rescher 1998: Ch. 2; Moscoso del Prado, ms.).
  2. How to compare various aspects of complexity to one another or their impact to global complexity?
  - How to compare rigid order and passive voice, or syllable complexity and theta role assignment?
  - More fundamentally: why should they be compared?

# 2.3 Types of complexity 1/2

- The notion of complexity needs to be broken down into types. For instance, Rescher (1998) proposes the following main “modes” of ontological complexity.
- Compositional complexity:
  - Constitutional (syntagmatic) complexity: the number of constituent elements (word length, sentence length).
  - Taxonomic (paradigmatic) complexity: the number of different types (e.g., inflectional paradigm, tense-aspect distinctions).
- Structural complexity:
  - Organizational complexity: the diversity of ways to arrange components in different types of interrelationship (e.g., variety of distinctive word orders).
  - Hierarchical complexity: elaborateness of subordination relationships (e.g., recursion).

# 2.3 Types of complexity 2/2

- The idea that we need different types of complexity has been presented in earlier linguistic work as well:
  - syntagmatic complexity, paradigmatic complexity, structural complexity, system complexity, conceptual complexity, structural elaboration, overspecification, economy, transparency, the principle of one-meaning–one-form, irregularity, hierarchical complexity.
- Each of these treats one of the following questions:
  - How to measure complexity at the surface structure level?
  - How to measure complexity at the level of semantic representation?
  - How about the mapping between form and meaning?

# 2.4 Description length

- What is the theoretical basis of linguistic complexity metrics? Or: how to connect them with the ways complexity is used in the sciences of complexity?
- Description length, or *Kolmogorov complexity*, is a common complexity metric in algorithmic information theory (Li and Vitányi 2008). It measures the length of the shortest description to specify a string, and it has often been used for measuring linguistic complexity (see Sinnemäki 2011).
- However, Kolmogorov complexity assigns high complexity values to random strings, which is counterintuitive (cf. Page 2012: 29–30).

# 2.4 Description length

00000000001111111111

→“Ten zeroes, followed by ten ones.” 6 words.

01001100011100001111

→“K zeroes then K ones: K from 1 to 4.” 9 words.

01101111010110010100

→ “Zero, two ones, zero, four ones, …”

- Takes at least 16 words to describe.
- Why not only 3 words: “It is random”. (Page 2012: 29–30).

# 2.4 Description length

- A solution is offered by the notion of *effective complexity* (Gell-Mann 1995). It measures the length of description required to specify *the set of regularities (or structure) in a string* rather than *the string itself* (cf. Dahl 2004: 24).
  - Compare Shakespeare's production with a description of the linguistic patterns that occur in Shakespeare's production.
- How effective complexity could be formalized is another story. I would argue that linguists' tools provide a good starting point, even crude ones.

# Overview of the talk

1. General background
2. A cross-linguistic approach
3. A case study
4. Conclusion

# 3. A case study

- It is often claimed that while languages vary in terms of local complexity, the differences are balanced out in cross-linguistic comparison (e.g., Hockett 1958: 180–181).
  → All languages are at about equal level of global complexity.

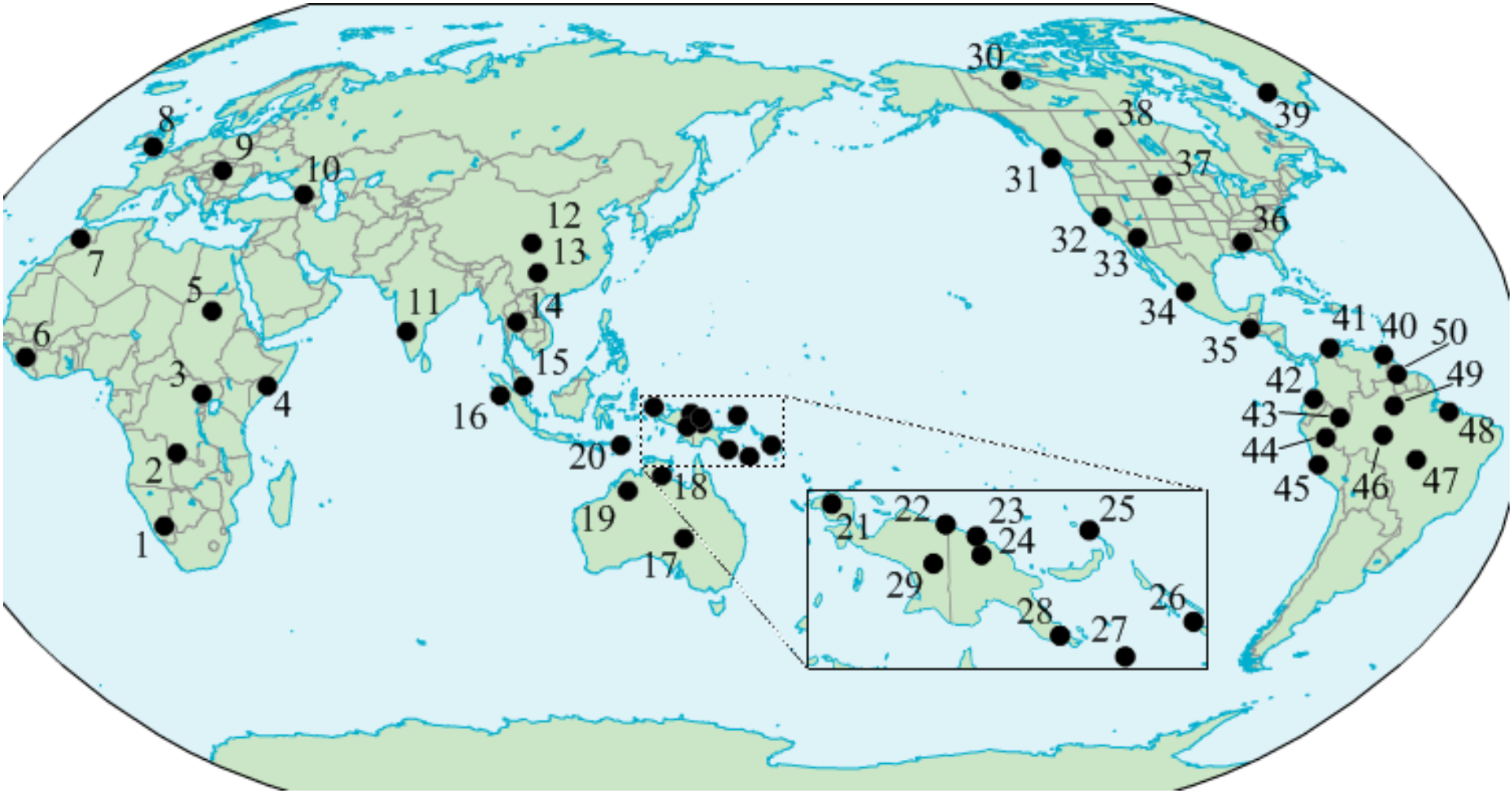
- Support for such trade-offs is often offered from the coding of syntactic relations: if a language has no case marking, it is likely to use rigid word order (e.g., Crystal 1997: 6).

- I present a cross-linguistic study on the coding of syntactic relations in a genealogically and areally stratified sample of 50 languages (Sinnemäki 2008).

# 3. The sample

- **Africa** (7 lgs): Khoekhoe [1], Lunda [2], Ngiti [3], Somali [4], Nubian (Dongolese) [5], Kisi [6], Berber (Middle Atlas) [7].
- **Eurasia** (4 lgs ): Welsh [8], Hungarian [9], Georgian [10], Kannada [11].
- **Southeast Asia-Oceania** (5 lgs): Qiang [12], Hmong Daw [13], Thai [14], Semelai [15], Nias [16].
- **Australia-New Guinea** (13 lgs): Diyari [17], Alawa [18], Gooniyandi [19], Klon [20], Maybrat [21], Skou [22], Arapesh [23], Yimas [24], Kuot [25], Lavukaleve [26], Yelî Dnye [27], Daga [28], Korowai [29].
- **North America** (10 lgs): Slave [30], Nuuchahnulth [31], Miwok (Southern Sierra) [32], Maricopa [33], Cora [34], Tzutujil [35], Choctaw [36], Lakhota [37], Cree (Plains) [38], Greenlandic (West) [39].
- **South America** (10 lgs): Warao [40], Ika [41], Quechua (Imbabura) [42], Yagua [43], Shipibo-Konibo [44], Jaqaru [45], Pirahã [46], Trumai [47], Urubú-Kaapor [48], Hixkaryana [49].
- **Creole** (1 lg): Berbice Dutch Creole [50].

# 3. Map of the sample languages



# 3. Argument coding

- The arguments of a transitive verb, here A(gent) and P(atient), can be coded via case marking or rigid word order (agreement excluded).
- In this study, case marking includes marking by inflectional and isolating formative as well as by tonal and morphophonological alternations (roughly dependent marking; Nichols 1992).
- Rigid word order occurs when a change in the order of the arguments triggers a change in the thematic interpretation of the sentence (Primus 1999: 132, 133), as in English:

  *the boy kissed the girl* vs. *the girl kissed the boy*.

  – Not the same as the degree of word order variation (Siewierska 1998).
- Only noun arguments treated here, pronouns excluded.

# 3. Presence of coding

- First, I noted whether case marking and/or rigid word order were used in each sample language. This measures zero vs. non-zero taxonomical complexity of case marking and organizational complexity of rigid word order.
  - Presence of a coding device requires longer description than its absence.
  - How come is rigid word order more complex than free word order? Overall probably not, but in a specific domain the presence of a constraint increases complexity.

Somer's Dxy = -.55; p < .001

x (predictor) = case marking

y = rigid order.

| | | Case marking | | |
|---|---|---|---|---|
| | | No | Yes | Total |
| Rigid Order | No | 3 | 18 | 21 |
| | Yes | 20 | 9 | 29 |
| | Total | 23 | 27 | 50 |

# 3. “Paradigm size” 1/2

- Second, I counted the number of different forms in the case paradigm (excluding morphophonological variation) and the number of different rigid orders.
- Ngiti (Kutsch Lojenga 1994: 193, 270) uses APV order in the present continuous and present habitual aspect and AVP order elsewhere → two rigid orders counted.

a. *Ma m-ɨ́ tsìtsì nɨ̄-ɔ̀nyʉ̀*.
1SG (A) 1SG-AUX banana (P) RSM-eat.NOM1 (V)
‘I am eating banana.’

b. *Nzonzo ɔ̀nyʉ̀ tsìtsì*.
children (A) eat.PFV.PRS (V) banana (P)
‘The children have eaten bananas.’

# 3. “Paradigm size” 2/2

- Diyari (Austin 1981: 48–49) has case marking for A and P. Eight slots in the paradigm and **five** different forms.

  *Ṯudu-yali puŋa yapi-ṇa wara-yi*.
  fire-ERG (A) hut-ABS (P) burn-PART (V) AUX-PRS
  ‘The fire burned the hut down.’ (Austin 1981: 118)

- Results: Dxy = -.59; p < .0001 (***).

| | erg | abs | acc |
|---|---|---|---|
| proper nouns | | | |
| • male personal names | *-li* | *-n̪a* | |
| • female personal names | *-ndu* | | *-n̪a* |
| common nouns | | | |
| • singular | *-yali* | *ø* | |
| • non-singular | *-li* | | *-n̪a* |

# 3. Number of constraints 1/2

- Third, I conducted a count of the constraints that are needed to specify case assignment or rigid order.
  - For instance, often only animate objects are case-marked while inanimate objects are zero-marked.
- In Diyari, the rules for assigning case in transitive clauses require three constraints:
  - individuation (proper/common), gender, and number.
- In Ngiti, the rules for rigid word order in transitive clauses require two constraints:
  - present continuous and present habitual.
- Results: Dxy = -.04; p = .77 (non-significant).
  - Why should a trade-off occur at the surface structure level rather than at the semantic level? Nubian, with pure accusative, has equal complexity in this regard than Kisi, which has no case marking.

# 3. Rule length

- Fourth, I tried to combine the paradigm size and the number of constraints by writing crude and experimenting rule descriptions for each coding device in each language and packed the rules with a zip-program. For instance:
  - Hixkaryana: “Rigid APV, used for fronting, rigid PVA elsewhere.”
  - Semelai: “A and P are optionally case-marked for disambiguation.”
  - Gooniyandi: “A is case-marked when human, optionally when non-human.”
- Results: Dxy = -.39; $p < .001$ (***).
  - Provides a slightly more comprehensive measure of complexity, but the result is still mostly affected by the presence of a coding device, not by the number of constraints.

# Overview of the talk

1. General background
2. A cross-linguistic approach
3. A case study
4. Conclusion

# 4. Conclusion

- Although complexity is difficult to formalize, at least a general-level characterization is possible.
- A cross-linguistic approach to grammatical complexity should be autonomous of the language user and focus on particular types of local complexity.
- There is cross-linguistic (statistical) evidence for a complexity trade-off between case marking and rigid word order, but limited in terms of type of complexity.
- Linguists' analytical tools, even crude ones, provide a feasible starting point for the measurement of grammatical complexity.

Thank you!

# References 1/2


Austin, P. 1981. *A grammar of Diyari, South Australia*. Cambridge: CUP.

Crystal, D. 1997. *The Cambridge encyclopedia of language* (2nd edn). Cambridge: CUP.

Dahl, Ö. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Edmonds, B. 1999. Syntactic measures of complexity. Ph.D. diss., University of Manchester.

Gell-Mann, M. 1995. What is complexity? *complexity* 1(1): 16-19.

Hawkins, J.A. 2004. *Efficiency and complexity in grammars*. Oxford: OUP.

Hockett, C.F. 1958. *A course in modern linguistics*. New York: Macmillan.

Hornby, A. & S. Wehmeier (compilers) 2007. *Oxford advanced learner's dictionary* (7th edn). Oxford: OUP.

Kusters, W. 2003. Linguistic complexity: The influence of social change on verbal inflection. Ph.D. diss., University of Leiden.

Kutsch Lojenga, C. 1994. *Ngiti: A Central-Sudanic language of Zaire*. Köln: Köppe.

Li, M. & P.M.B. Vitányi 2008. *An introduction to Kolmogorov complexity and its applications* (3rd edn). New York: Springer.

Lloyd, S. 2001. Measures of complexity: A nonexhaustive list. *IEEE Control Systems Magazine* 21(4): 7-8.

Mikulecky, Don C. 2001. The emergence of complexity: Science coming of age or science growing old. *Computers & Chemistry* 25: 341-348.

# References 2/2


Miestamo, M. 2008. Grammatical complexity in a cross-linguistic perspective. In M. Miestamo et al. (eds.), Language complexity: Typology, contact, change, 23-41. Amsterdam: John Benjamins.

Moscoso del Prado Martín, F. 2010. The effective complexity of language: English requires at least an infinite grammar. Manuscript.

Nichols, J. 1992. *Linguistic diversity in space and time*. Chicago: The University of Chicago Press.

Page, S.E. 2011. *Diversity and complexity*. Princeton: Princeton University Press.

Primus, B. 1999. *Cases and thematic roles: Ergative, accusative and active*. Tübingen: Niemeyer.

Rescher, N. 1998. *Complexity: A philosophical overview*. New Brunswick: Transaction.

Simon, H.A. 1996. *The sciences of the artificial* (3rd edn). Cambridge, MA: MIT Press.

Sinnemäki, K. 2008. Complexity trade-offs in core argument marking. In M. Miestamo et al. (eds.), *Language complexity: Typology, contact, change*, 67-88. Amsterdam: John Benjamins.

Sinnemäki, K. 2011. Language universals and linguistic complexity: Three case studies in core argument marking. Ph.D. dissertation, University of Helsinki. Available at https://helda.helsinki.fi/handle/10138/27782.