

Språk och Datorer (729G49)

Språkteknologi 5: Korpusutveckling & Interbedömarreliabilitet

12/5/2025



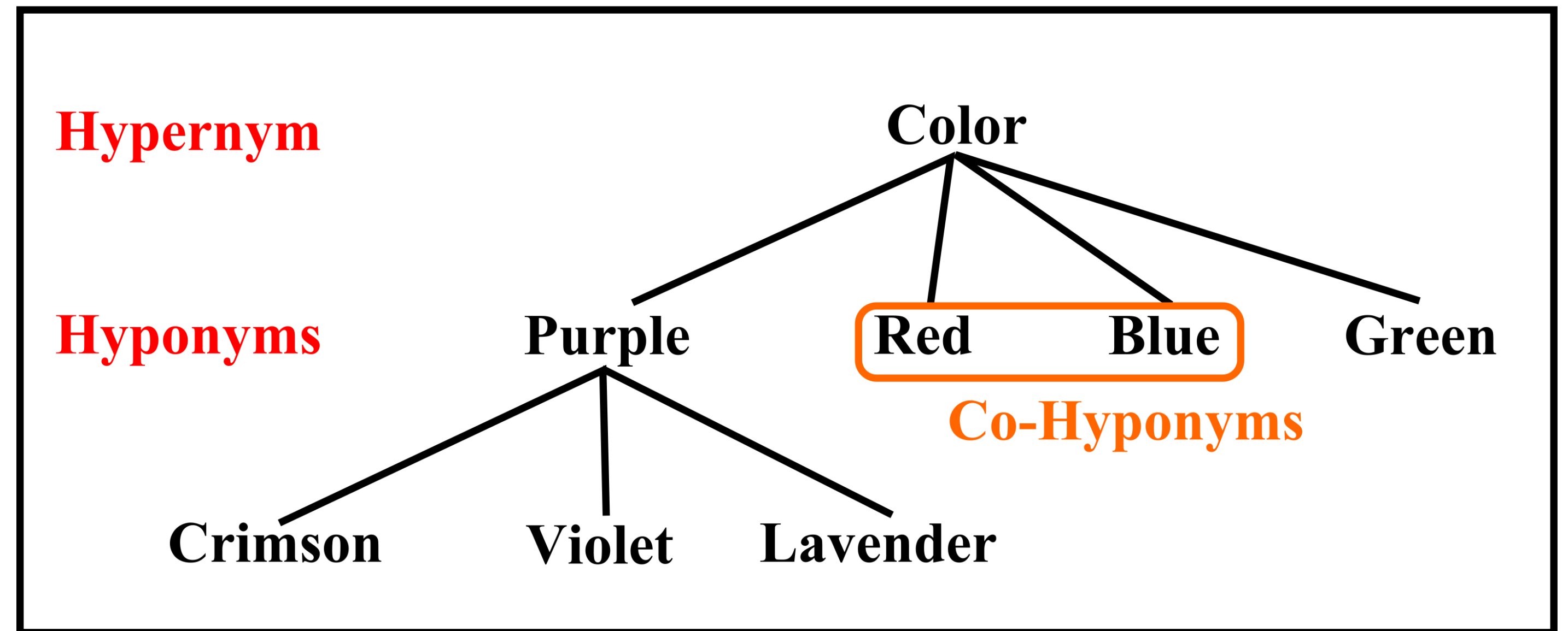
Innehåll

- Hyponymer & hyperonymer.
- Korpusutveckling.
- Interbedömarreliabilitet & Cohen's Kappa.

Hyponymmer och hyperonymer

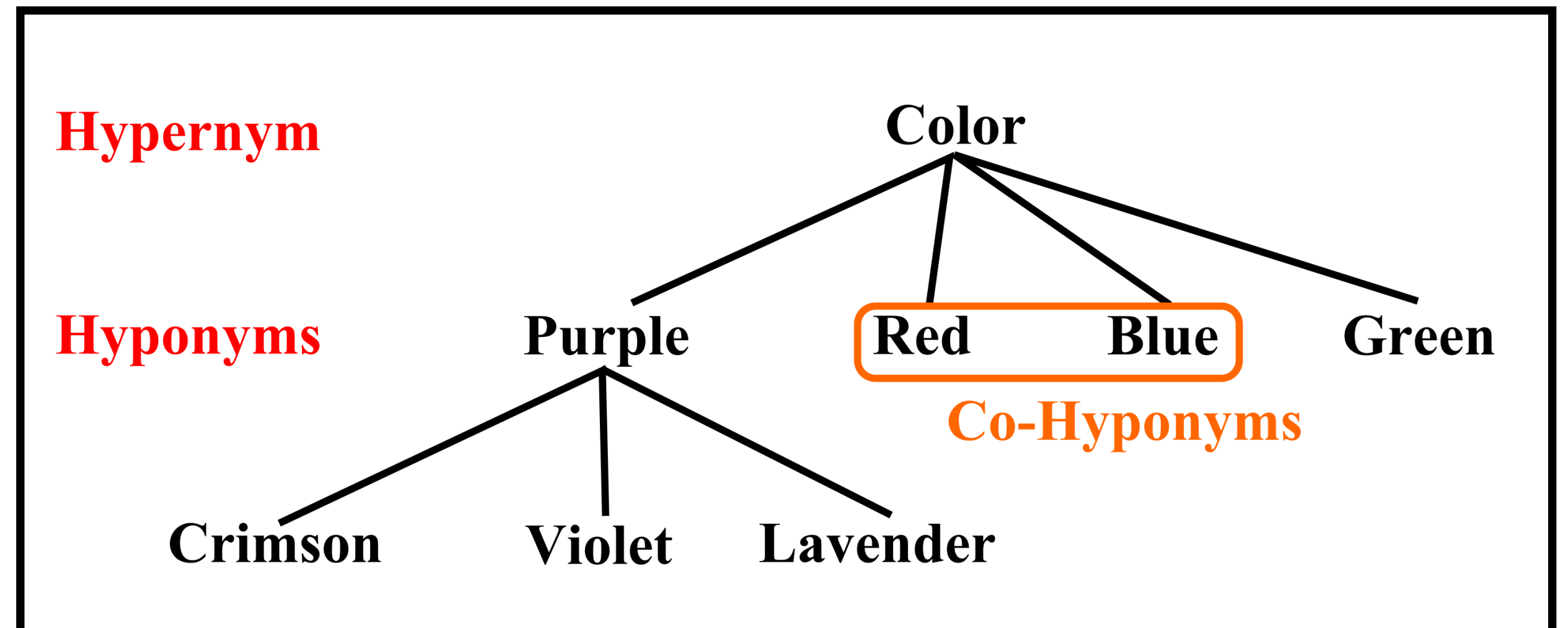
Hyponymer och Hyperonymer (1)

- En **hyponym** är ett ord vars betydelse innefattas av ett annat ords betydelse; detta andra ord kallas **hyperonym**.
- En hyponym är **mindre abstrakt** och innehåller **mer information** än sin hyperonym.



Hyponymer och Hyperonymer (2)

- Hyponymi kan uttryckas som "är en/ett"-relation:
 - *Röd är en färg.*
 - *En bil är ett fordon.*
 - *En mango är en frukt.*
 - *Sverige är ett land.*
 - *Linköping är en stad.*



Hyperonymer i Översättningar



- *Det är min farmors bil.*
- *This is my paternal grandma's car.*

Hyperonymer i Översättningar



- *Det är min farmors bil.*
- *This is my ~~paternal~~ grandma's car.*

Undvik översättningsengelska :)

WordNet (1)



- En stor databas för engelska med ordbetydelser och semantiska relationer.
- En ordbetydelse representeras som en mängd av ömsesidigt synonyma ord, ett så kallat synset.
- Den centrala semantiska relationen mellan ordbetydelser är hyponymi-hyperonymi.
- Webbversion: <https://en-word.net/>

WordNet (2)

Nouns

(n) hyponym, subordinate, subordinate word *a word that is more specific than a given word*

Hypernyms (1)

(n) word *a unit of language that native speakers can identify "words are the blocks from which sentences are made" "he hardly said ten words all morning"*

Hypernyms (1)

(n) language unit, linguistic unit *one of the natural units into which linguistic messages can be analyzed*

Hypernyms (1)

(n) component, component part, constituent, part, portion *something determined in relation to something that includes it "he wanted to feel a part of something bigger than himself" "I read a portion of the manuscript" "the smaller component is hard to reach" "the animal constituent of plankton"*

Hypernyms (1)

(n) relation *an abstraction belonging to or characteristic of two entities or parts together*

Hypernyms (1)

(n) abstract entity, abstraction *a general concept formed by extracting common features from specific examples*

Hypernyms (1)

(n) entity *that which is perceived or known or inferred to have its own distinct existence (living or nonliving)*

WordNet (3)

Nouns

(n) word *a unit of language that native speakers can identify "words are the blocks from which sentences are made" "he hardly said ten words all morning"*

Hypernyms (1)

Hyponyms (56)

(n) anagram *a word or phrase spelled by rearranging the letters of another word or phrase*

MORE ►

(n) anaphor *a word (such as a pronoun) used to avoid repetition; the referent of an anaphor is determined by its antecedent*

MORE ►

(n) antonym, opposite, opposite word *a word that expresses a meaning opposed to the meaning of another word, in which case the two words are antonyms of each other "to him the antonym of 'gay' was 'depressed'"*

MORE ►

(n) paronym *a word that strongly resembles another word in spelling*

MORE ►

(n) back-formation *a word invented (usually unwittingly by subtracting an affix) on the assumption that a familiar word derives from it*

MORE ►

(n) charade *a word acted out in an episode of the game of charades*

MORE ►

Från <https://en-word.net/>

WordNet (3)

Nouns

(n) word *a unit of language that native speakers can identify "words are the blocks from which sentences are made" "he hardly said ten words all morning"*

Hypernyms (1)

Hyponyms (56)

(n) anagram *a word or phrase spelled by rearranging the letters of another word or phrase*

MORE ►

(n) anaphor *a word (such as a pronoun) used to avoid repetition; the referent of an anaphor is determined by its antecedent*

MORE ►

(n) antonym, opposite, opposite word *a word that expresses a meaning opposed to the meaning of another word, in which case the two words are antonyms of each other "to him the antonym of 'gay' was 'depressed'"*

MORE ►

(n) paronym *a word that strongly resembles another word in spelling*

MORE ►

(n) back-formation *a word invented (usually unwittingly by subtracting an affix) on the assumption that a familiar word derives from it*

MORE ►

(n) charade *a word acted out in an episode of the game of charades*

MORE ►

Från <https://en-word.net/>

Hearst-Mönster (1)

- Artikel: Hearst (1992)
- En metod för att automatiskt extrahera hyponymer från lingvistiskt annoterad text.

Automatic Acquisition of Hyponyms from Large Text Corpora

Marti A. Hearst
Computer Science Division, 571 Evans Hall
University of California, Berkeley
Berkeley, CA 94720
and
Xerox Palo Alto Research Center
marti@cs.berkeley.edu

Abstract

We describe a method for the automatic acquisition of the hyponymy lexical relation from unrestricted text. Two goals motivate the approach: (i) avoidance of the need for pre-encoded knowledge and (ii) applicability across a wide range of text. We identify a set of lexico-syntactic patterns that are easily recognizable, that occur frequently and across text genre boundaries, and that indisputably indicate the lexical relation of interest. We describe a method for discovering these patterns and suggest that other lexical relations will also be acquirable in this way. A subset of the acquisition algorithm is implemented and the results are used to augment and critique the structure of a large hand-built thesaurus. Extensions and applications to areas such as information retrieval are suggested.

1 Introduction

Currently there is much interest in the automatic acquisition of lexical syntax and semantics, with the goal of building up large lexicons for natural lan-

guage processing.

following sentence:¹.
(S1) The bow lute, such as the Bambara ndang, is plucked and has an individual curved neck for each string.

Most fluent readers of English who have never before encountered the term “Bambara ndang” will nevertheless from this sentence infer that a “Bambara ndang” is a kind of “bow lute”. This is true even if the reader has only a fuzzy conception of what a bow lute is. Note that the author of the sentence is not deliberately defining the term, as would a dictionary or a children’s book containing a didactic sentence like *A Bambara ndang is a kind of bow lute*. However, the semantics of the lexico-syntactic construction indicated by the pattern:

(1a) NP_0 such as $\{NP_1, NP_2 \dots, (and \mid or)\} NP_n$

are such that they imply

(1b) for all NP_i , $1 \leq i \leq n$, $hyponym(NP_i, NP_0)$

Thus from sentence (S1) we conclude

Hearst-Mönster (2)

Mönster	Exempel
and other	...temples, tresuries, and other civic buildings.
or other	Bruises, wounds, broken bones or other injuries ...
such as	The bow lute, such as the Bambara ndang
Such.... as	...such authors as Herrick, Goldsmith, and Shakespeare.
including	... common-law countries, including Canada.
, especially	European countries, especially France and Spain,...

Korpusutveckling

Korpusutveckling

- Korpusutvecklingen består av flera steg, blant annat:
 - Planering.
 - Förberedning av datan.
 - Annotering.
 - Kvalitetskontroll och utvärdering.

Planera

- Syftesdefinition: Vad ska korpusen användas till?
- Urval av material:
 - Vilka texttyper ska ingå? (t.ex. tidningsartiklar, talad dialog, sociala medier)
 - Hur representativ ska korpusen vara?
- Val av korpusformat (t.ex. CoNLL, XML, JSON).

Förbereda data

- Insamling av data
- Rensning och normalisering av data:
 - Ta bort oönskat material (t.ex. annonser, dubbletter).
 - Avformatera.
 - Standardisera format (t.ex. filformat, kodning som UTF-8).
 - I vissa fall även stavningsnormalisering.

Annotera

- Lingvistisk annotering: t.ex.
 - Tokenisering
 - Lemmatisering
 - Part-of-speech (POS)
 - Syntaktisk struktur
- Metadata-annotering: uppgifter om t.ex. författare, publikation, datum.

Kvalitetskontroll

- Dokumentation av alla steg: urvalskriterier, annoteringsprinciper, tekniska specifikationer.
- Kontroll av annoteringsnoggrannhet:
 - Manuell granskning.
 - Interbedömarreliabilitet.

Kvalitetskontroll och Utvärdering

- Dokumentation av alla steg: urvalskriterier, annoteringsprinciper, tekniska specifikationer.
- Kontroll av annoteringsnoggrannhet:
 - Manuell granskning.
 - Interbedömarreliabilitet.
- Andra mått av hur bra korpusen uppfyller förväntningarna (beroende av korpusens syfte).

Interbedömarreliabilitet

Reliabilitet

- Reliabilitet syftar på hur pålitlig en kvantitativ undersökning är, hur mycket vi kan tro på resultaten från den.
 - Påverkas resultaten av de som genomfört undersökningen?
 - Påverkas resultaten av tiden?
 - Finns det samstämmighet mellan olika delresultat?

Interbedömarreliabilitet

- För att pröva hur mycket en undersökning påverkas av de som genomfört den kan man mäta interbedömarereliabilitet.
- Man låter undersökningen genomföras av flera olika bedömare och mäter överensstämmelsen mellan dem.
 - Olika mått beroende på typ av undersökning.
- Det naturliga komplementet till interbedömarereliabilitet är intrabedömarereliabilitet, som fokuserar på en bedömare.
 - Hur mycket påverkas bedömningarna av dagsformen?



Made with <https://gemini.google.com/> ; Prompt: generate a manga image of two linguistic annotators who heavily disagree with each other in front of a computer

Cohen's Kappa (1)

- Cohens kappa (κ) mäter överensstämmelse mellan två bedömare vid en **klassifikationsuppgift**.
 - Hyponym | Ingen hyponym
- Måttet ställer den relativa **observerade överensstämmelsen** mot sannolikheten för **slumpmässig överensstämmelse**.

Cohen's Kappa (2)

- Beräkning: $\kappa \equiv \frac{p_o - p_e}{1 - p_e}$
- p_o : observerad överensstämmelse
- p_e : förväntad överensstämmelse

Cohen's Kappa (3)

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e}$$

Den grad av överensstämmelse som inte kan förklaras av slumpen

- p_o : observerad överensstämmelse
- p_e : förväntad överensstämmelse

Cohen's Kappa (4)

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e}$$

Den grad av överensstämmelse som inte kan förklaras av slumpen

Den maximal möjliga grad av överensstämmelse som inte kan förklaras av slumpen

- p_o : observerad överensstämmelse
- p_e : förväntad överensstämmelse

Cohen's Kappa (5)

- Beräkning: $\kappa \equiv \frac{p_o - p_e}{1 - p_e}$
- p_o : observerad överensstämmelse
- p_e : förväntad överensstämmelse
- Värdet $\kappa = 1$ anger fullkomlig överensstämmelse.
- Vid fullkomlig avsaknad av överensstämmelse är $\kappa = 0$.

Observerad överensstämmelse (1)

	B godkänd	B underkänd
A godkänd	20	5
A underkänd	10	15

Observerad överensstämmelse (2)

- I hur många procent av fallen överensstämmer bedömarna med varandra?
- För att beräkna den observerade överensstämmelsen summerar vi längs diagonalen och delar den med det totala antalet instanser i tabellen.

	B godkänd	B underkänd
A godkänd	20	5
A underkänd	10	15



Observerad överensstämmelse (2)

- Total antal: 50.
- Samma: 20 + 15
- Observerad överensstämmelse:

$$\frac{20 + 15}{50} = 0.7$$

	B godkänd	B underkänd
A godkänd	20	5
A underkänd	10	15



Slumpmässig överensstämmelse (1)

- Två bedömare som slumpar fram ja/nej-svar kommer att överensstämma i 50% av fallen.
- Men den förväntade slumpmässiga överensstämmelsen beror på **antalet** och **fördelningen** av klasserna.
- Överensstämmelse mellan bedömare är intressant endast om den avviker från (helst mycket högre än) den förväntade slumpmässiga överensstämmelsen.

Slumpmässig överensstämmelse (2)

- A ger godkänd 25 gånger.
- Total antal: 50.

	B godkänd	B underkänd
A godkänd	20	5
A underkänd	10	15

- $$\frac{20 + 5}{50}$$

Slumpmässig överensstämmelse (3)

- B ger godkänd 30 gånger.
- Total antal: 50.

	B godkänd	B underkänd
A godkänd	20	5
A underkänd	10	15

- $$\frac{20 + 10}{50}$$

Slumpmässig överensstämmelse (4)

- Förväntad överensstämmelse för *godkänd* är alltså:

- $\frac{25}{50} * \frac{30}{50} = 0.3$

- På samma sätt beräknar vi den förväntade överensstämmelse för *underkänd*:

- $\frac{25}{50} * \frac{20}{50} = 0.2$

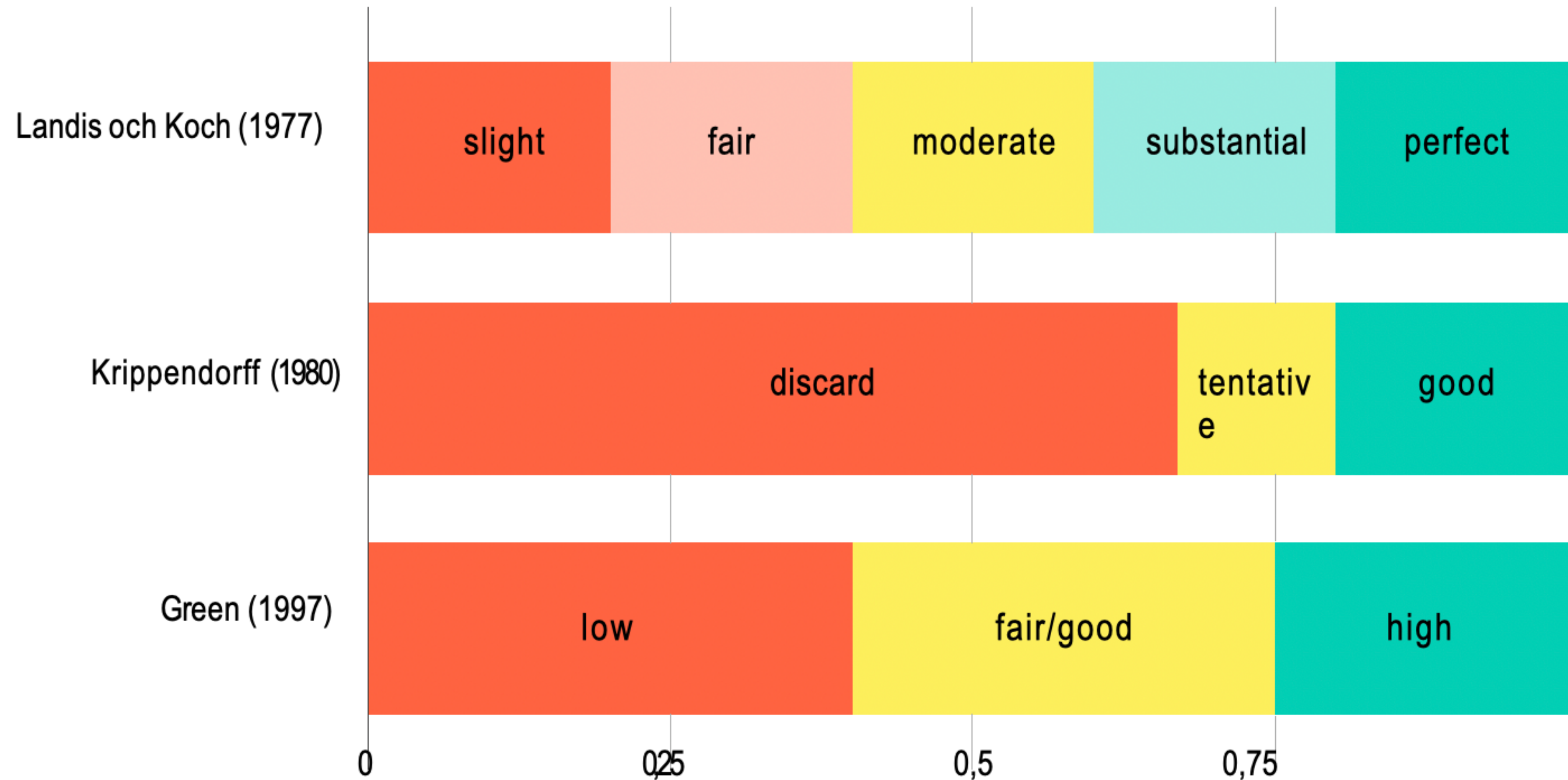
- $P_e = 0.3 + 0.3 = 0.5$

Cohen's Kappa

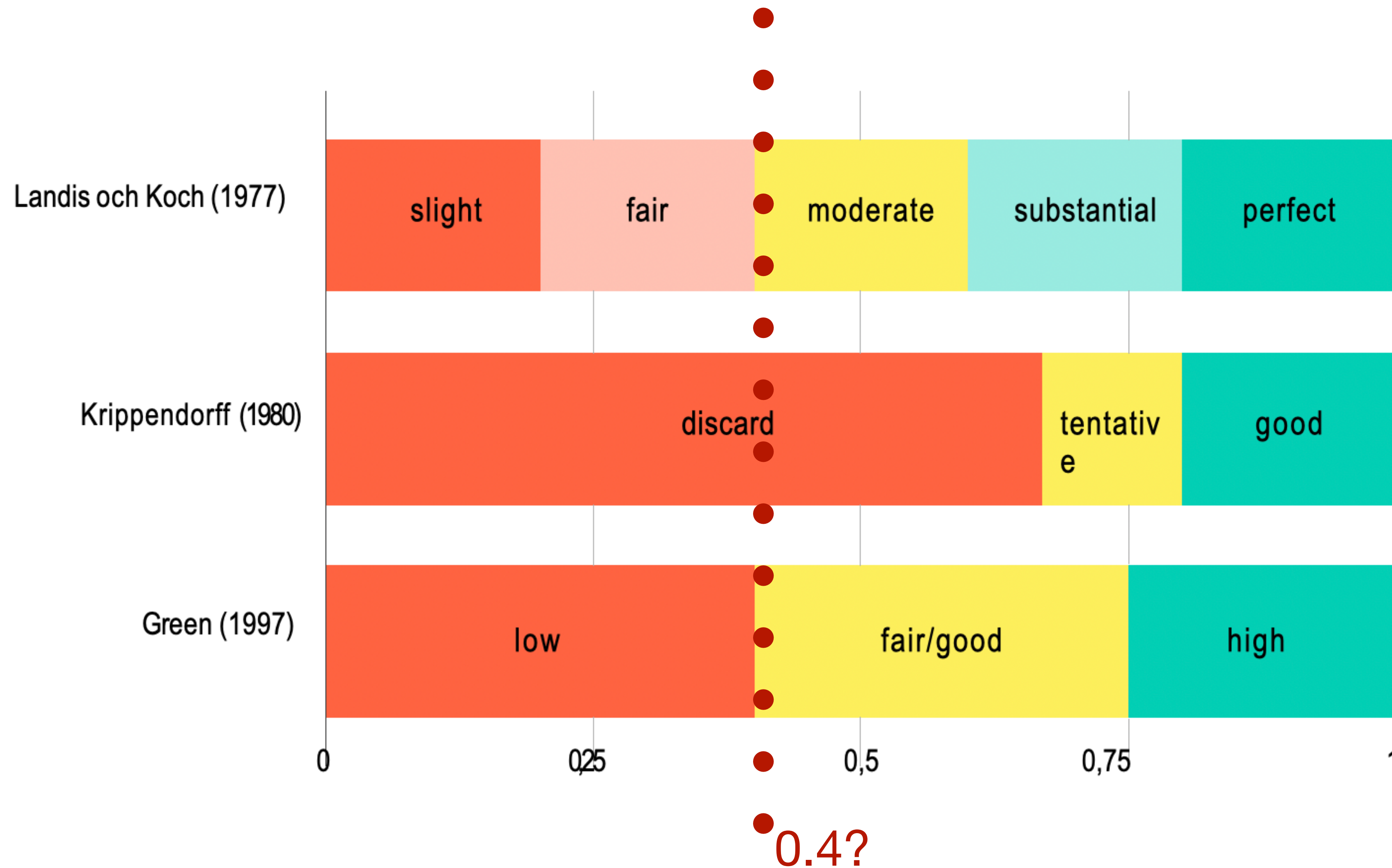
- $\kappa \equiv \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$

- Vad betyder det?

Tolkning av Cohen's Kappa



Tolkning av Cohen's Kappa



L5

Tack för idag!