Reference databases, Research databases, and Typological databases

Jeff Good, University at Buffalo (jcgood@buffalo.edu)

NYU Workshop on the Feasibility of a Web-based Database of the Syntactic Structures of the World's Languages

What is a database?

- Reference database: A database containing data whose structure is presumed to be well understood that is used to facilitate access to information in some knowledge domain
- Research database: A database containing data whose structure is poorly understood that used to facilitate research on the nature of the knowledge domain from which the data is drawn

Some examples

Some examples

- An online lexicon is usually employed as a reference database

Some examples

- An online lexicon is usually employed as a reference database
- WALS, in its present form, is a reference database

Some examples

- An online lexicon is usually employed as a reference database
- WALS, in its present form, is a reference database
- Many typological databases start out as research databases but finish as reference databases

Some examples

- An online lexicon is usually employed as a reference database
- WALS, in its present form, is a reference database
- Many typological databases start out as research databases but finish as reference databases
- Autotyp is designed to be a hybrid reference/research database

Difficulties

Difficulties

- Research databases present a number of difficulties

Difficulties

- Research databases present a number of difficulties
- It can be quite difficult to build a database without a stable model

Difficulties

- Research databases present a number of difficulties
- It can be quite difficult to build a database without a stable model
- There is no "end" in terms of technological development—results prompt revisions and new needs

Difficulties

- Research databases present a number of difficulties
- It can be quite difficult to build a database without a stable model
- There is no "end" in terms of technological development—results prompt revisions and new needs
- Community consensus on data entry essentially impossible

This project

This project

- “Syntax” is, obviously, not well understood

This project

- “Syntax” is, obviously, not well understood
- Some kinds of syntactic data are understood well enough to allow for consensus structures

This project

- "Syntax" is, obviously, not well understood
- Some kinds of syntactic data are understood well enough to allow for consensus structures
- I wouldn't say that syntactic "features" and "values" are among these

WALS

WALS

- WALS is an instructive example

WALS

- WALS is an instructive example
- It is not one database, but rather 141 databases using a common metadata standard

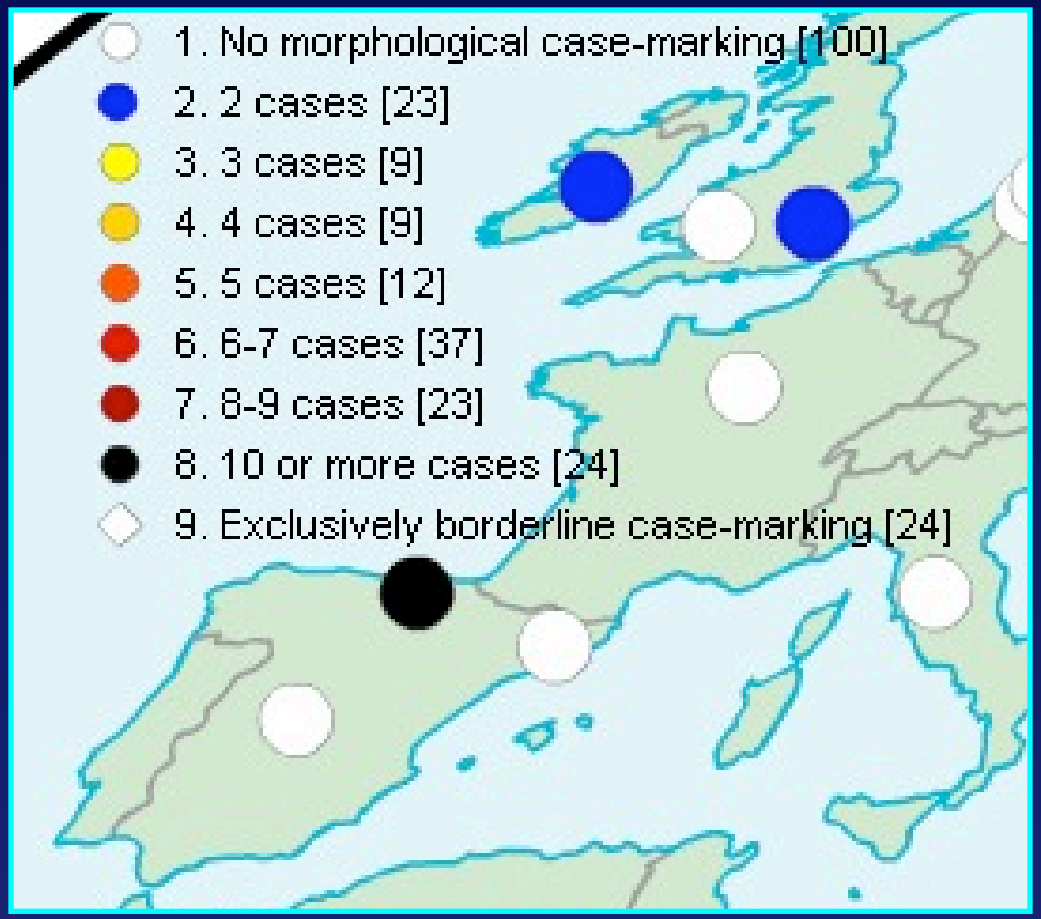
WALS

- WALS is an instructive example
- It is not one database, but rather 141 databases using a common metadata standard
- Datapoints between databases cannot assumed to be directly comparable

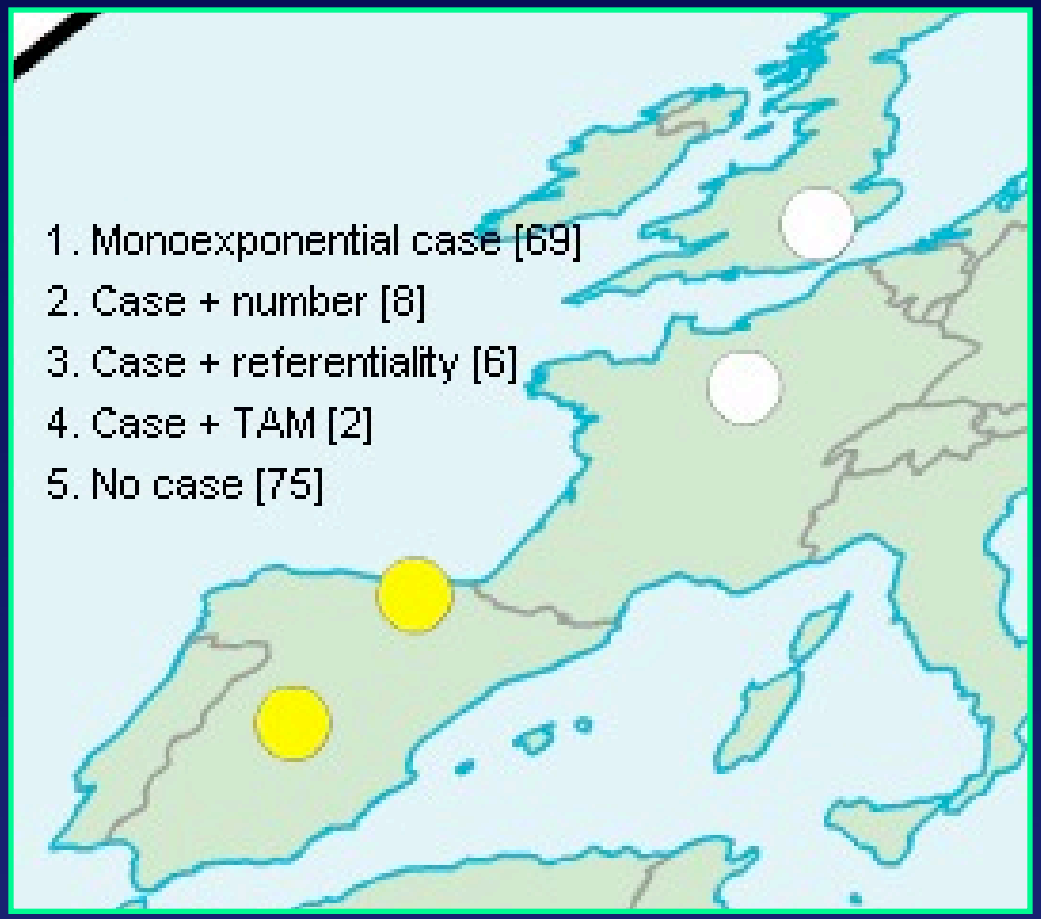
WALS

- WALS is an instructive example
- It is not one database, but rather 141 databases using a common metadata standard
- Datapoints between databases cannot assumed to be directly comparable
- Databases are to be cited as independent works

How many cases?



"Number of Cases"
by Oliver A. Iggesen



"Exponence of Selected Grammatical Formatives"
by Balthasar Bickel and Johanna Nichols

(See also: http://email.eva.mpg.de/~cysouw/pdf/cysouwSOCIALLAYER.pdf)

How many cases?

2. 2 cases [23]

5. No case [75]

"Number of Cases"
by Oliver A. Iggesen

"Exponence of Selected Grammatical Formatives"
by Balthasar Bickel and Johanna Nichols

(See also: http://email.eva.mpg.de/~cysouw/pdf/cysouwSOCIALLAYER.pdf)

Which is “correct”

Which is “correct”

- So, which is correct? Does English have two cases or none?

Which is “correct”

- So, which is correct? Does English have two cases or none?
- Of course, it depends what you mean by “case”.

Which is “correct”

- So, which is correct? Does English have two cases or none?
- Of course, it depends what you mean by “case”.
- A possible solution: Give a clear, strict definition of *case*

Which is “correct”

- So, which is correct? Does English have two cases or none?
- Of course, it depends what you mean by “case”.
- A possible solution: Give a clear, strict definition of *case*
- But, this is a reference solution, not a research solution

Which is "correct"

- So, which is correct? Does English have two cases or none?
- Of course, it depends what you mean by "case".
- A possible solution: Give a clear, strict definition of *case*
- But, this is a reference solution, not a research solution
- Different definitions may be both reasonable and interesting

(My) starting principles

(My) starting principles

- We understand the data, but we don't understand the typology

(My) starting principles

- We understand the data, but we don't understand the typology
- Therefore, the database must be minimally restrictive from a typological standpoint

(My) starting principles

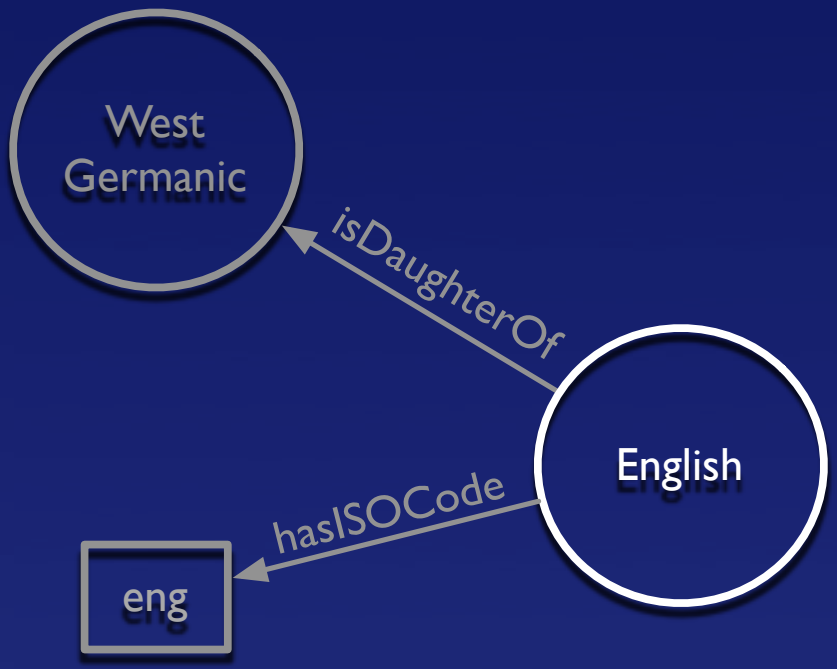
- We understand the data, but we don't understand the typology
- Therefore, the database must be minimally restrictive from a typological standpoint
- The problem is not to develop a typological model but to support:

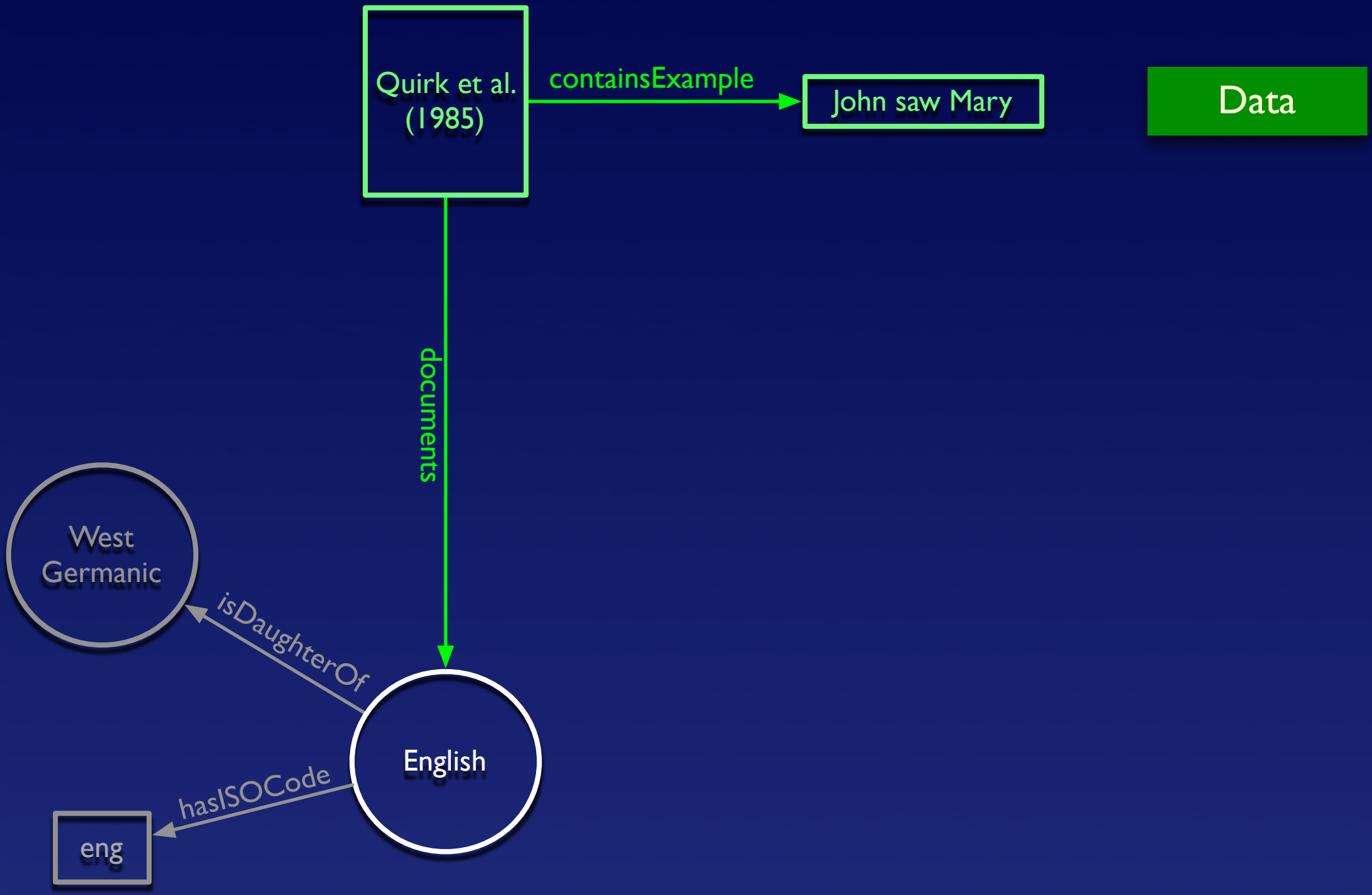
(My) starting principles

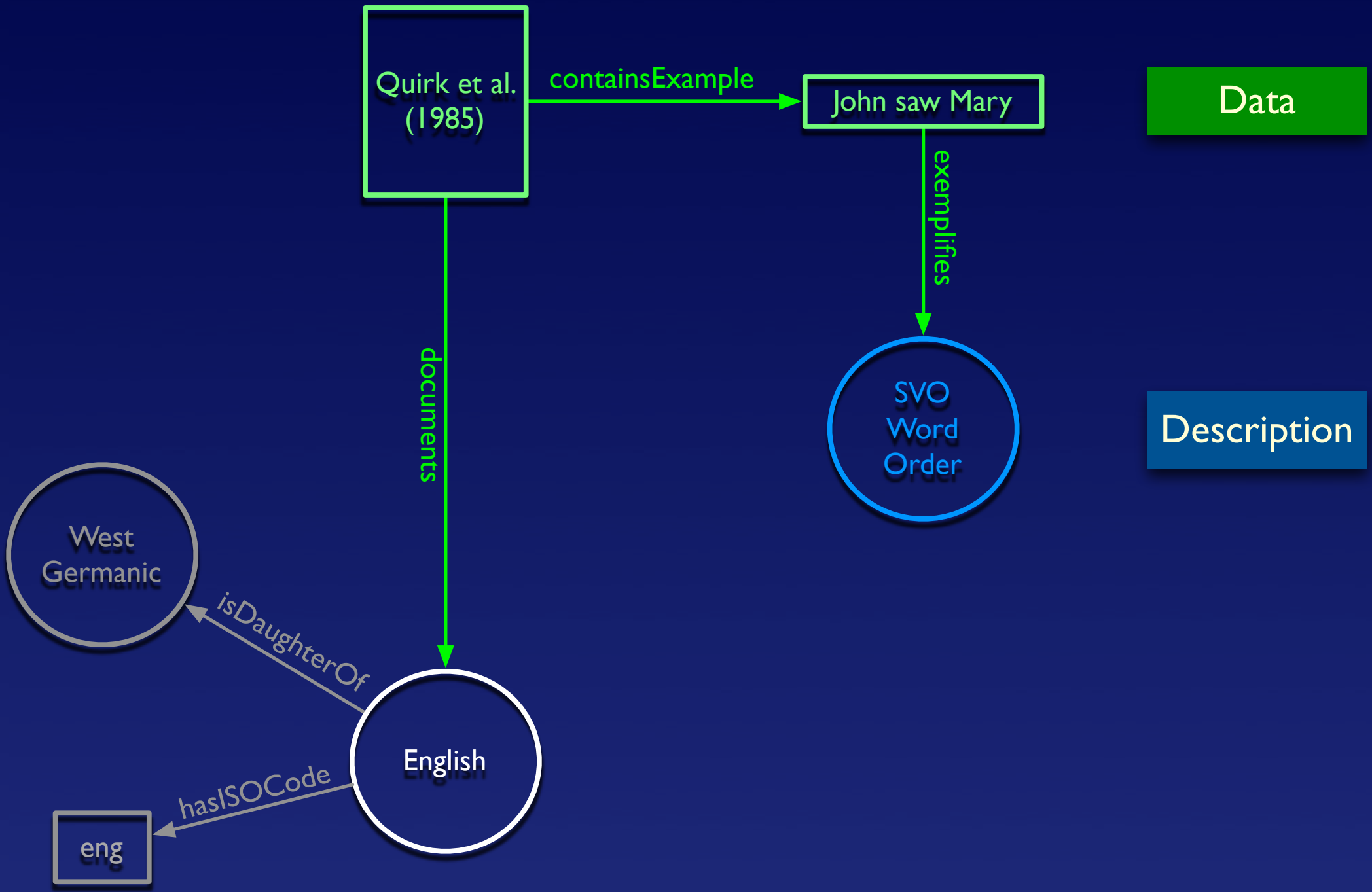
- We understand the data, but we don't understand the typology
- Therefore, the database must be minimally restrictive from a typological standpoint
- The problem is not to develop a typological model but to support:
  - Independent research projects in typology

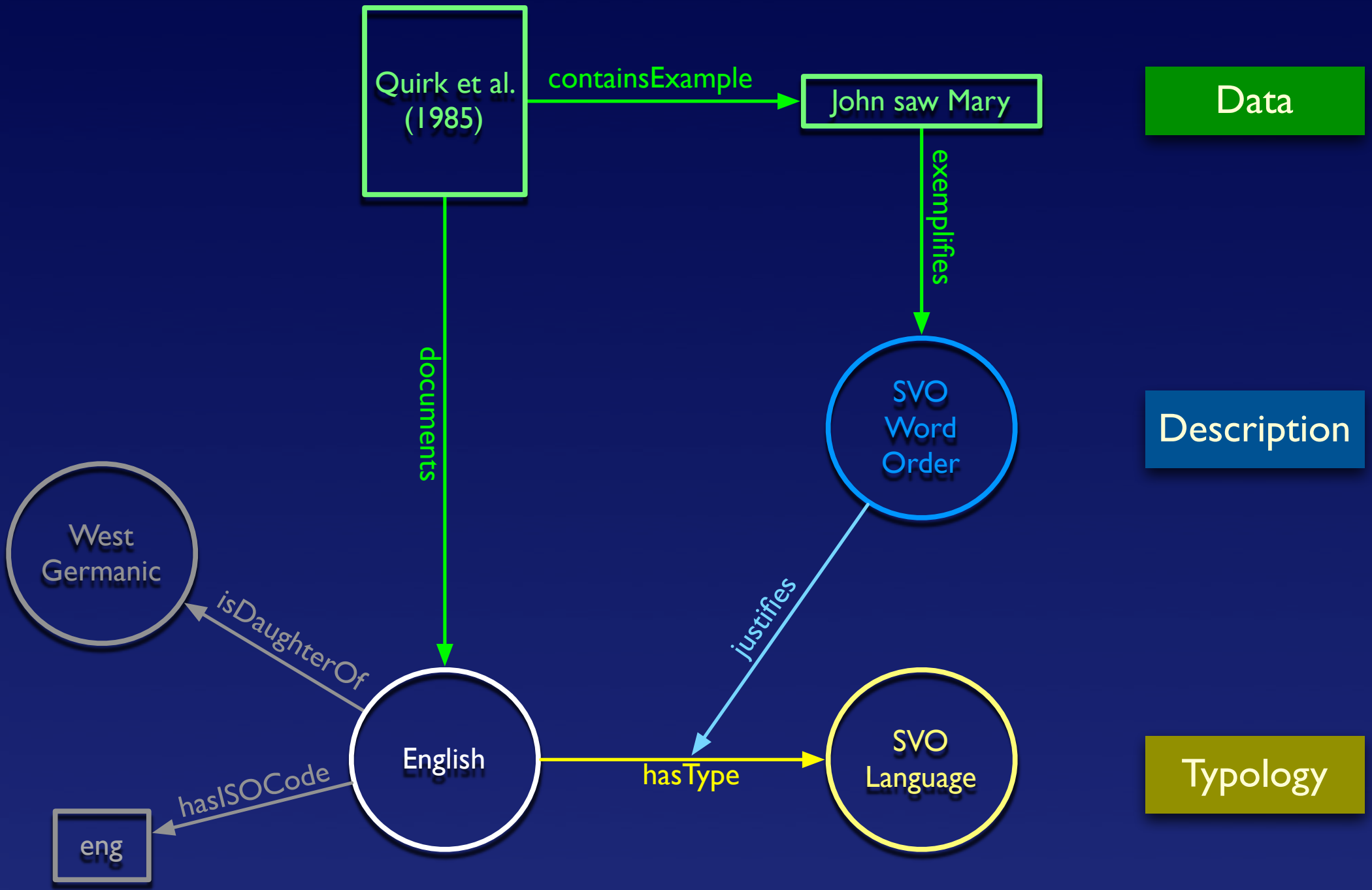
(My) starting principles

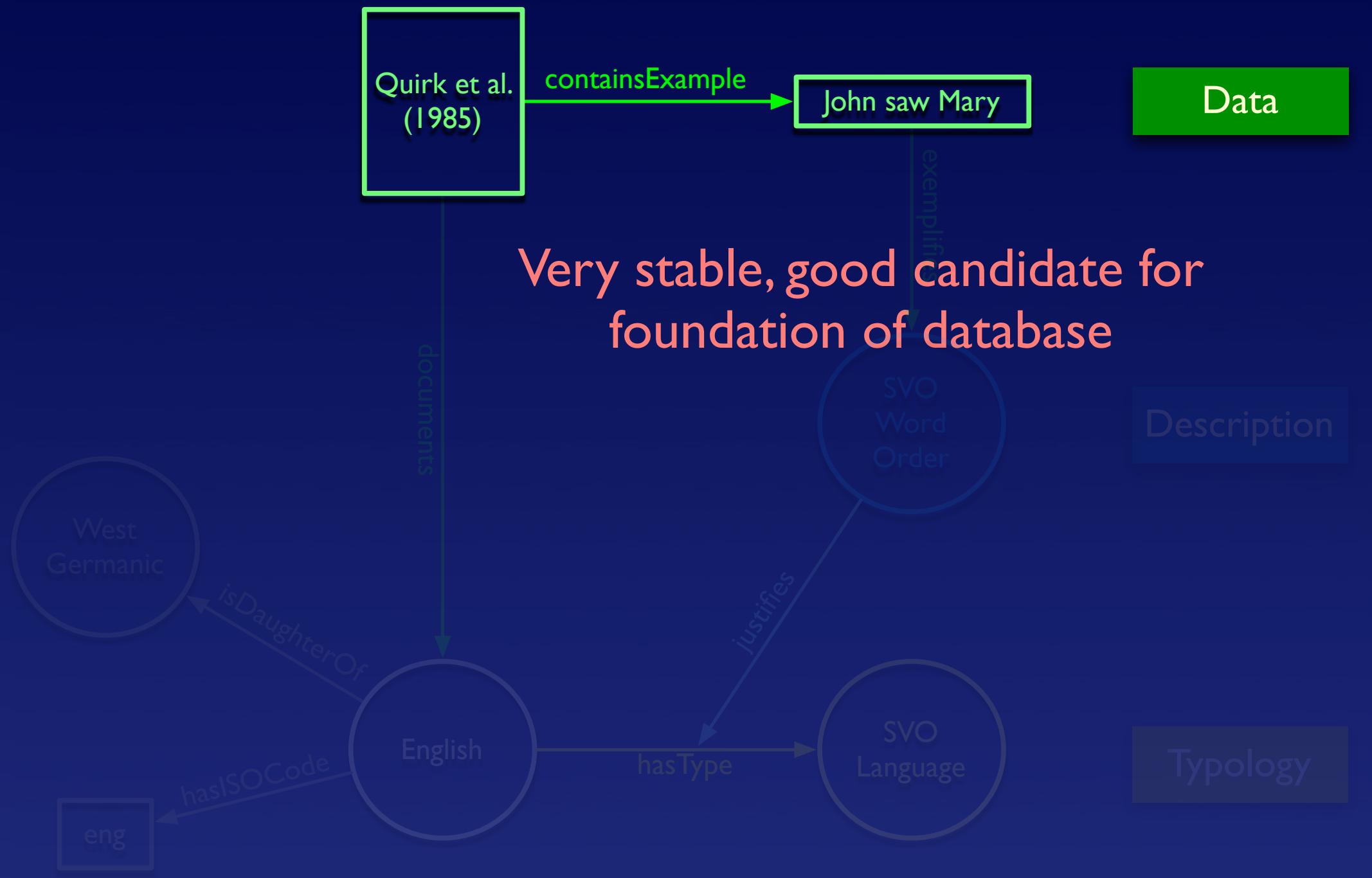
- We understand the data, but we don't understand the typology
- Therefore, the database must be minimally restrictive from a typological standpoint
- The problem is not to develop a typological model but to support:
  - Independent research projects in typology
  - A reasonable level of interoperability among those research projects


West
Germanic
isDaughterOf
English
hasISOCode
eng

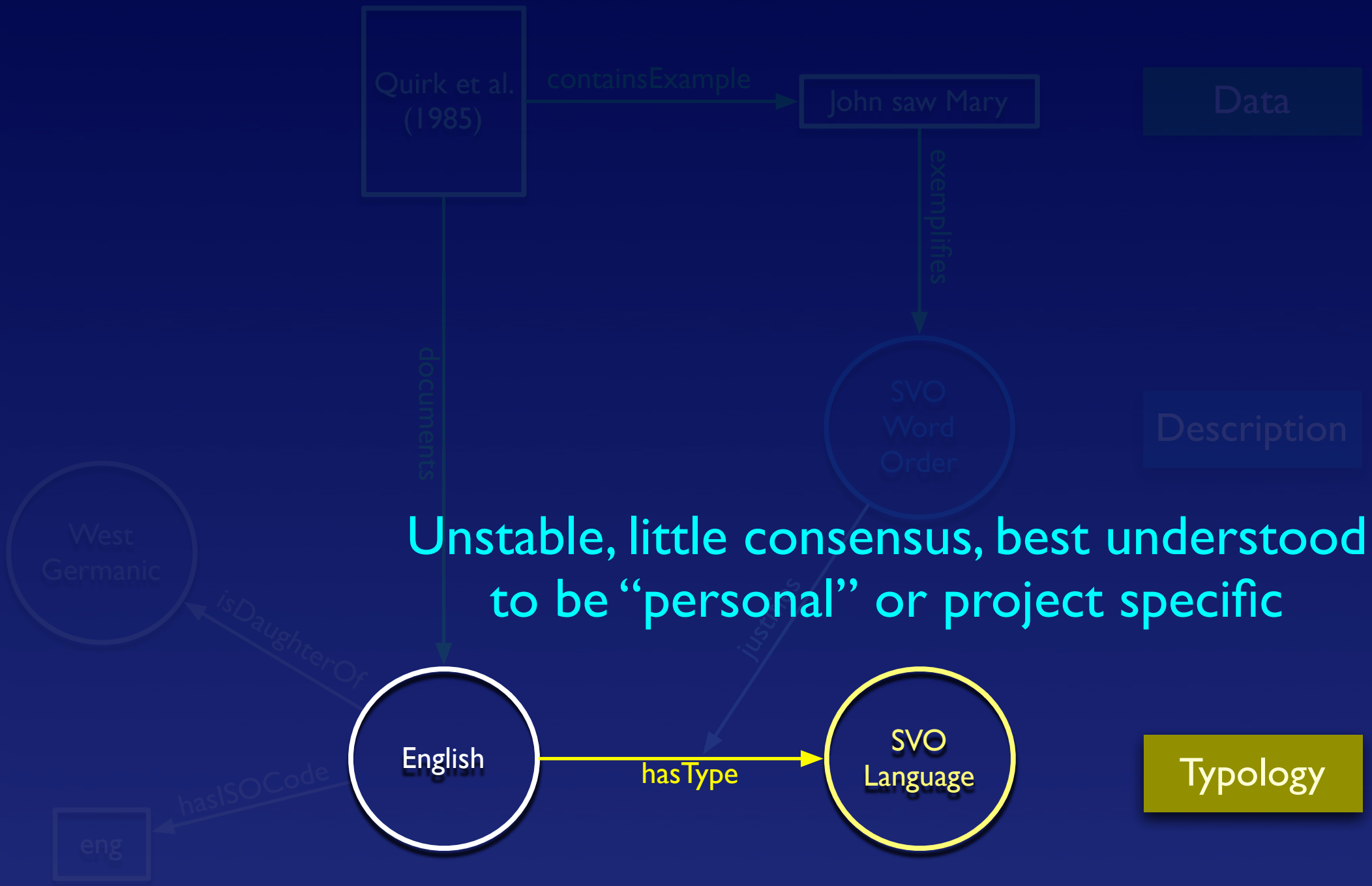

Quirk et al. (1985)
containsExample
John saw Mary
Data
documents
West Germanic
isDaughterOf
English
hasISOCode
eng


Quirk et al. (1985)
containsExample
John saw Mary
exemplifies
documents
SVO Word Order
West Germanic
isDaughterOf
English
hasISOCode
eng
Data
Description

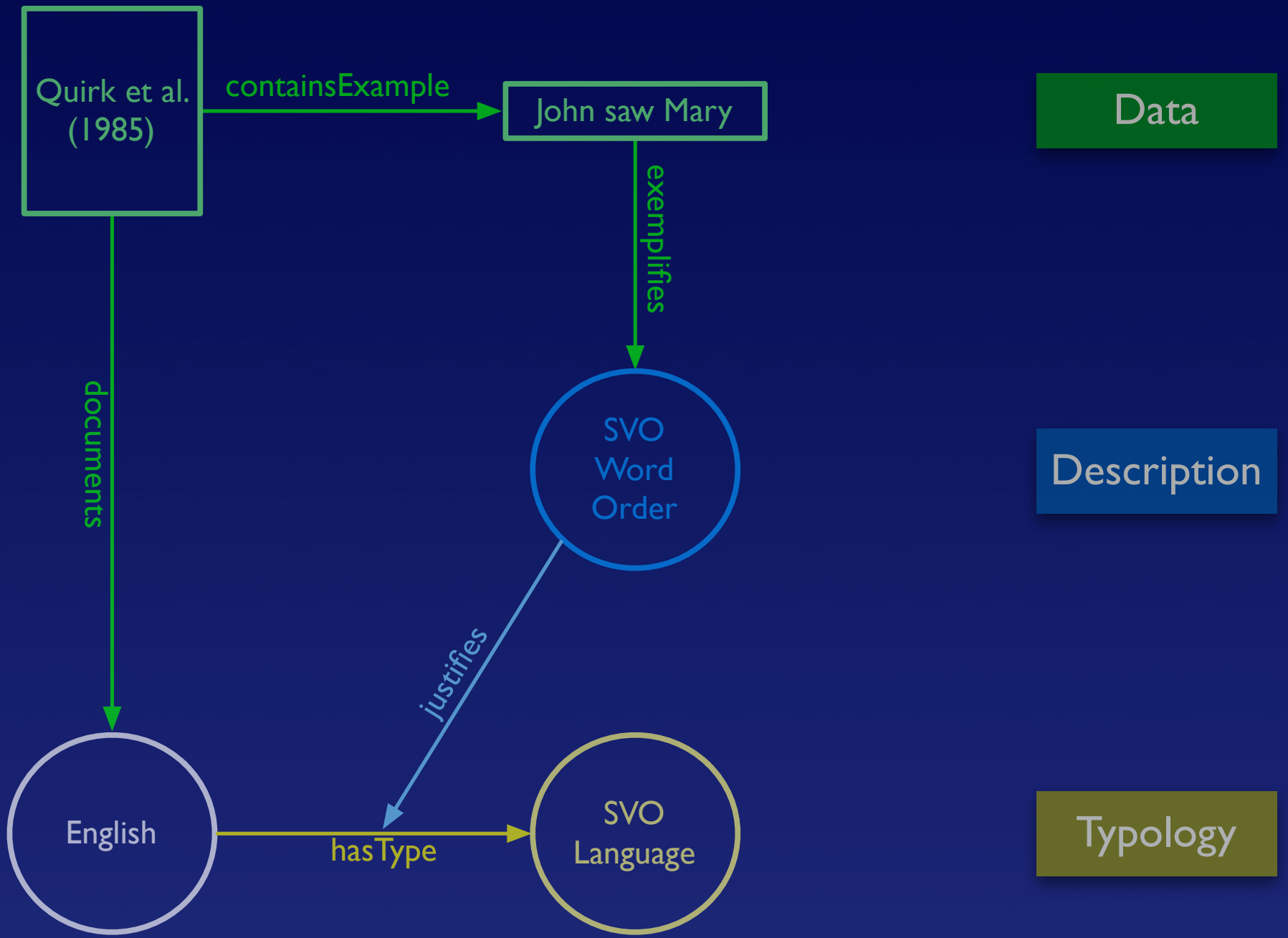

Quirk et al. (1985)
containsExample
John saw Mary
exemplifies
documents
SVO Word Order
West Germanic
isDaughterOf
justifies
English
hasISOCode
hasType
SVO Language
eng
Data
Description
Typology

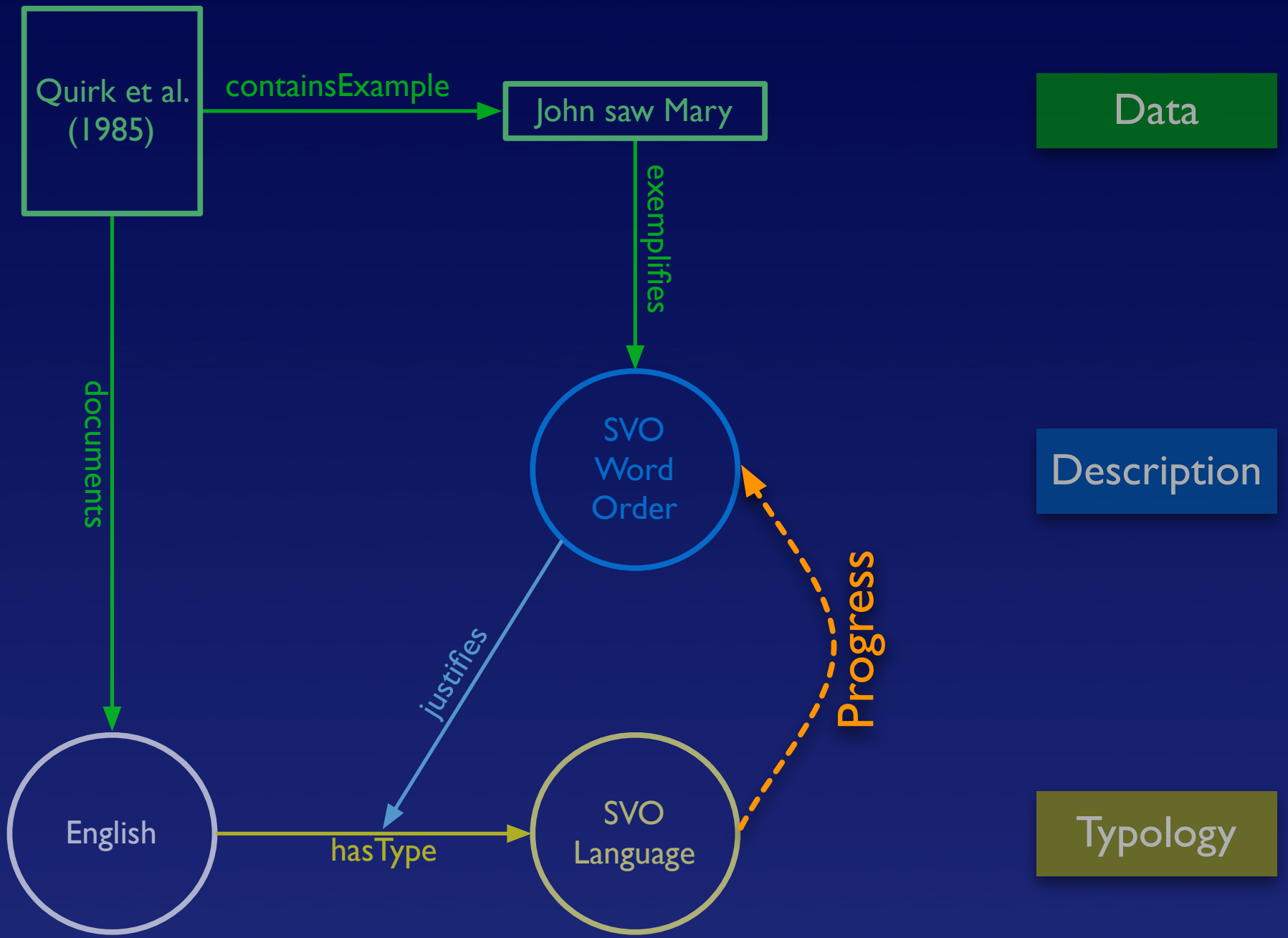

Quirk et al. (1985)
containsExample
John saw Mary
Data
exemplifies
Very stable, good candidate for foundation of database
documents
SVO Word Order
Description
West Germanic
isDaughterOf
justifies
English
hasType
SVO Language
Typology
hasISOCode
eng


Quirk et al. (1985)
containsExample
John saw Mary
exemplifies
Data
Less stable, but consensus probably achievable
documents
SVO Word Order
Description
West Germanic
isDaughterOf
justifies
English
hasType
SVO Language
Typology
hasISOCode
eng



Unstable, little consensus, best understood to be "personal" or project specific


Quirk et al. (1985)
containsExample
John saw Mary
exemplifies
documents
SVO Word Order
justifies
English
hasType
SVO Language
Data
Description
Typology


Quirk et al. (1985)
containsExample
John saw Mary
exemplifies
documents
SVO Word Order
justifies
Progress
English
hasType
SVO Language
Data
Description
Typology

Recommendation

- To build a typological research database
  - Attempt some level of standardization for data and description
  - Do not standardize typology
  - Provide tools for typologization based on description
- Reference databases like WALS should be built on top of such a system, not into it