

Automatically Learning Topical Content

Sean Boisen, [myfirstname]@logos.com

Director of Content Innovation, Logos Bible Software

BibleTech:2013, Seattle WA March 15, 2013

Presentation URL: <http://semanticbible.org/other/talks/2013/BibleTech/LCV.html>

Outline

- The Logos Controlled Vocabulary (LCV)
-

Spoke previously about the LCV here at BibleTech 2010. Today i'll focus on how we're mining it for data.

- The LCV corpus
- Learning from the corpus
 - Topic importance
 - References
 - Key terms
 - Topic relationships

Semantic Organization at Logos

Logos ... organizing the world's Biblical information

- the Bible Knowledgebase: people, places, things and events described in the Biblical text

Released Logos 5 last November: using italics for new items since last BibleTech

- the Logos Controlled Vocabulary
- *Referent Analysis*
- *Bible Sense Lexicon*
- *Timeline*
- *Preaching Themes*

The Logos Controlled Vocabulary (LCV)

- LCV = a *controlled vocabulary* for Biblical studies
 - **controlled vocabulary**: an organized system for labeling content using words and phrases, to improve search and navigation
 - Supports topic search and exploration, connects other knowledge
 - Mitigates problems of ambiguity, homographs, synonyms, variation in spelling and other kinds of representation
- Provides *semantic organization* of concepts

Not keywords
- Organizes content from Bible dictionaries, encyclopedias, and other subject-oriented secondary resources, regardless of how the publisher expressed it.

LCV as a database of conceptual knowledge is distinct from its use in annotating texts: most of today's talk is about this latter use

LCV Example: Deceit

COULD USE A BETTER EXAMPLE HERE. Fill out with other relationships

This organizes knowledge about language and subjects in biblical studies

- Attributes:

Type: lexical	Name: False
EN Preferred label: "Deceit"	Extra-biblical: False
EN Alternate labels: "Deception", "Deceitful", "Deceive", "Connivance", ...	Compound: False
Importance: 0.4737	Breadth: 0.67

- Relationships:

Related concepts: none
Preaching Theme: Honesty

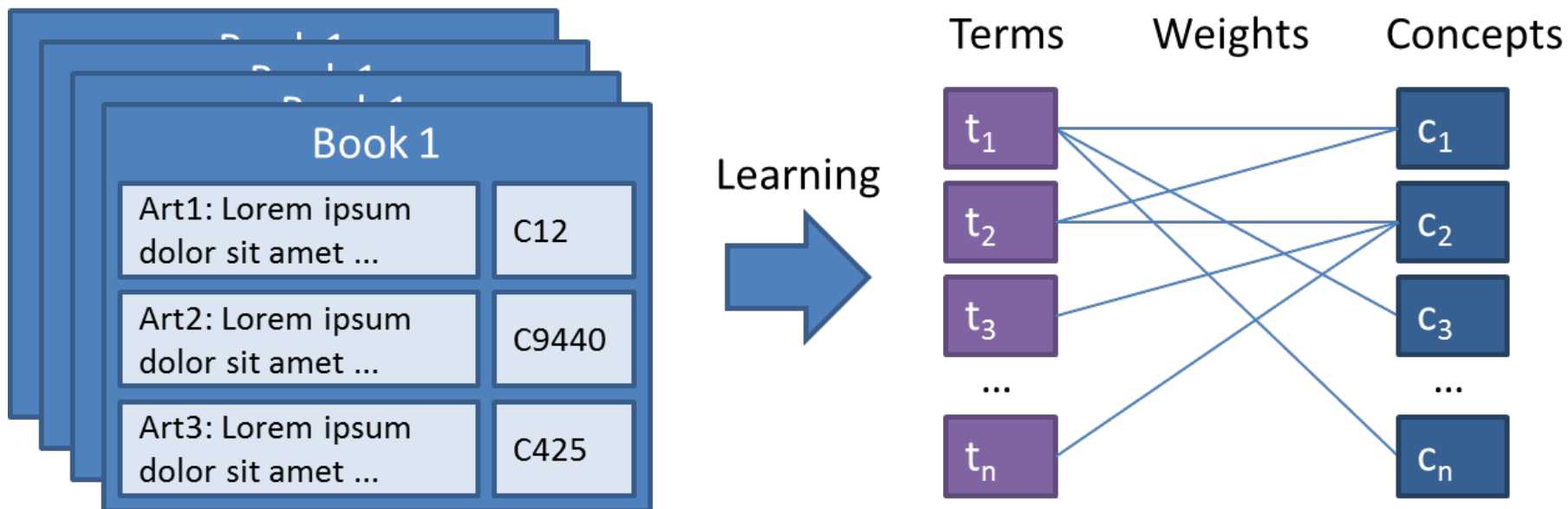
Logos Topics: <http://topics.logos.com/Deceit>

Newer LCV Features

Connecting different datasets

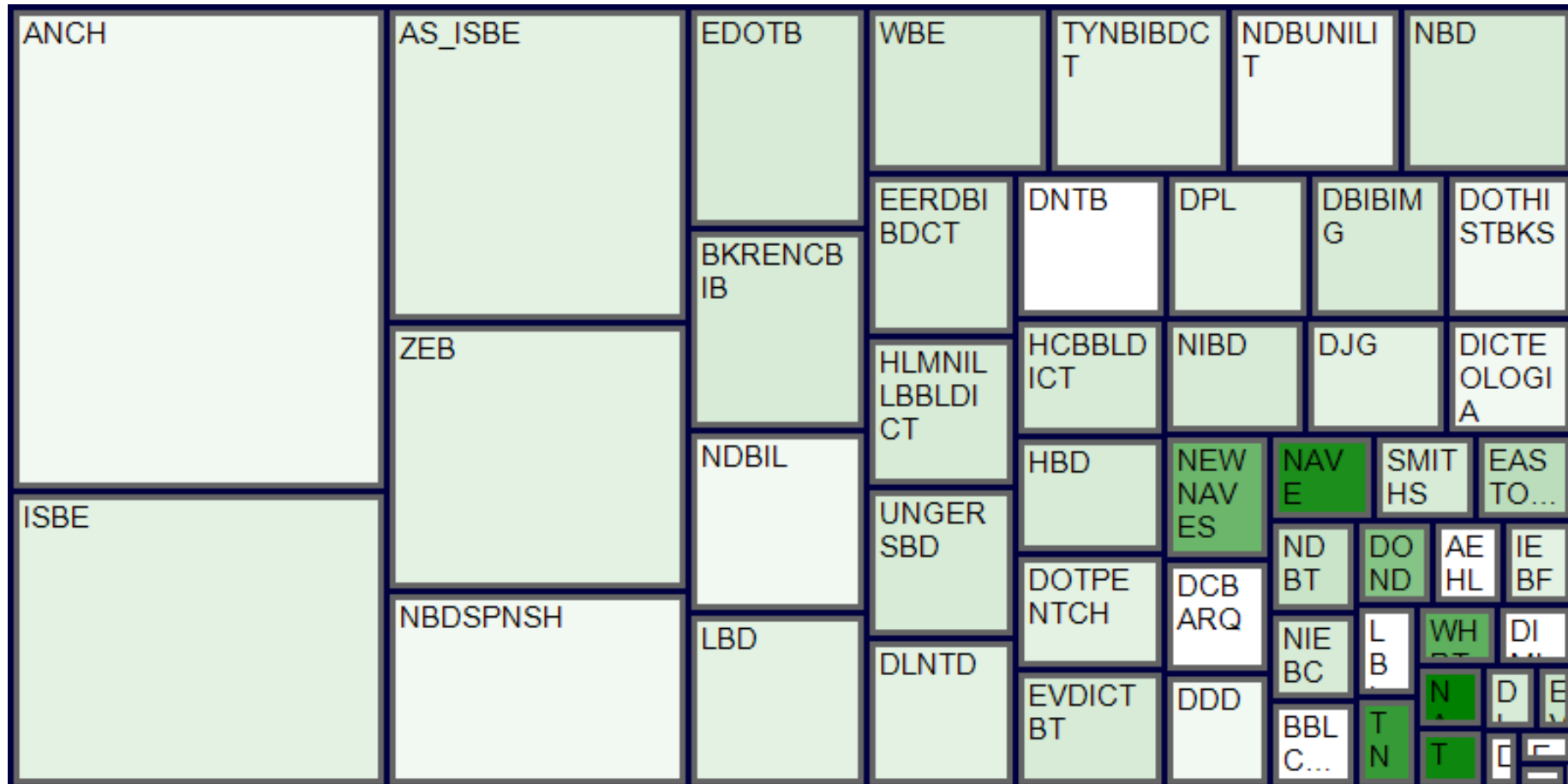
- Incorporated Spanish terms and resources
- Added hierarchy (broader/narrower concepts)
- Linked to Preaching Themes, Bible Sense Lexicon
- Disambiguated compound concepts for ambiguous names
- Added Topic Guide to expose more LCV features
- Aligned more resources with the LCV

Connecting Concepts to Terms and References



Concept	An item from the LCV (a "topic")
Book	Traditional container for content
Article	Subject-oriented unit of prose from a book
Term	A "word" from an article
Reference	A Bible reference term

The LCV Corpus: Overview



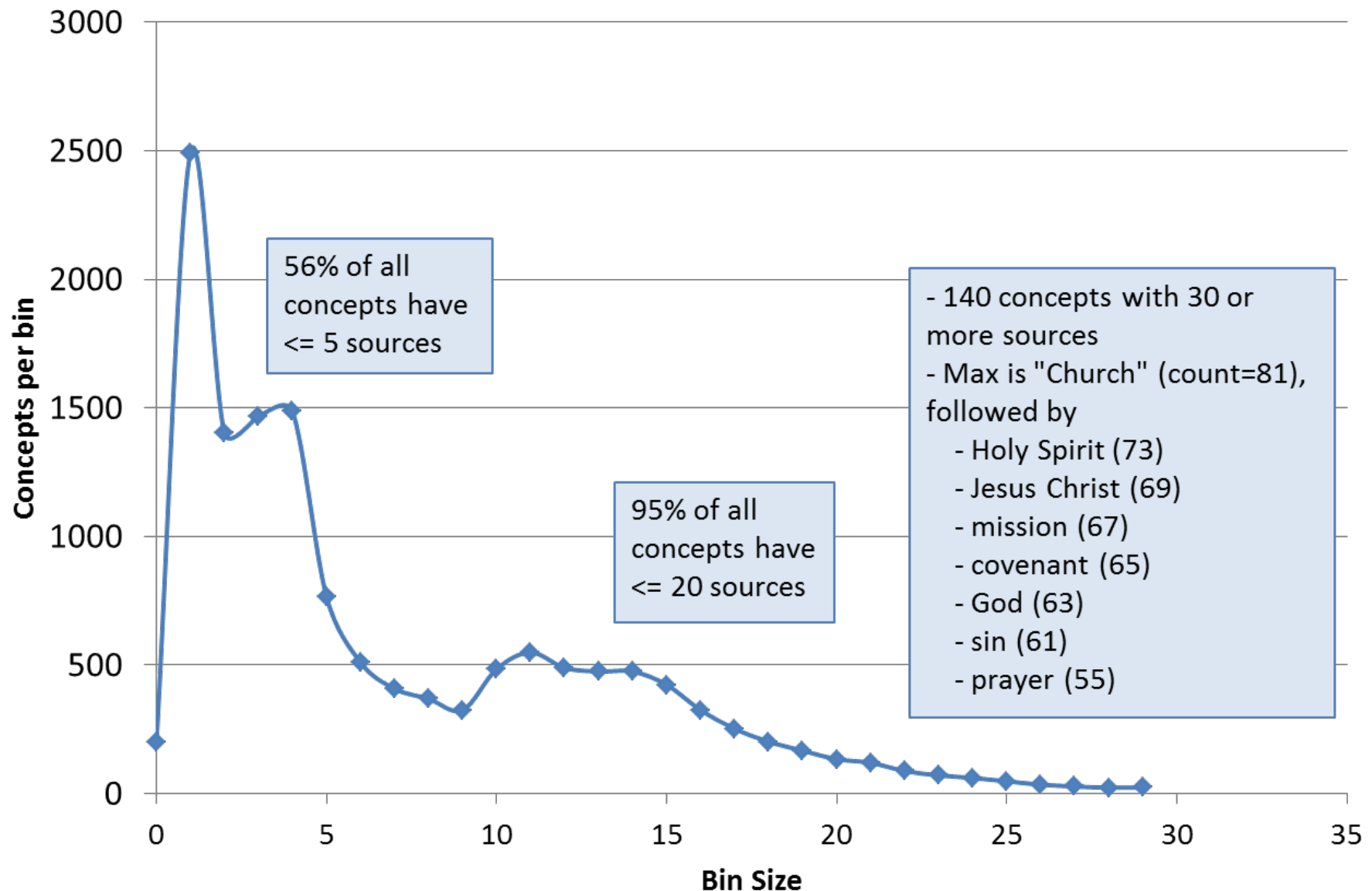
Here i've combined books in a series: elsewhere in the counts, they're considered separate books

- 140k articles (34M words) annotated with concepts
- ~50 "independent" sources

- Varying styles and purposes

Breadth of Sources for Concepts

Learning Importance: with 13k concepts, need a way to rank



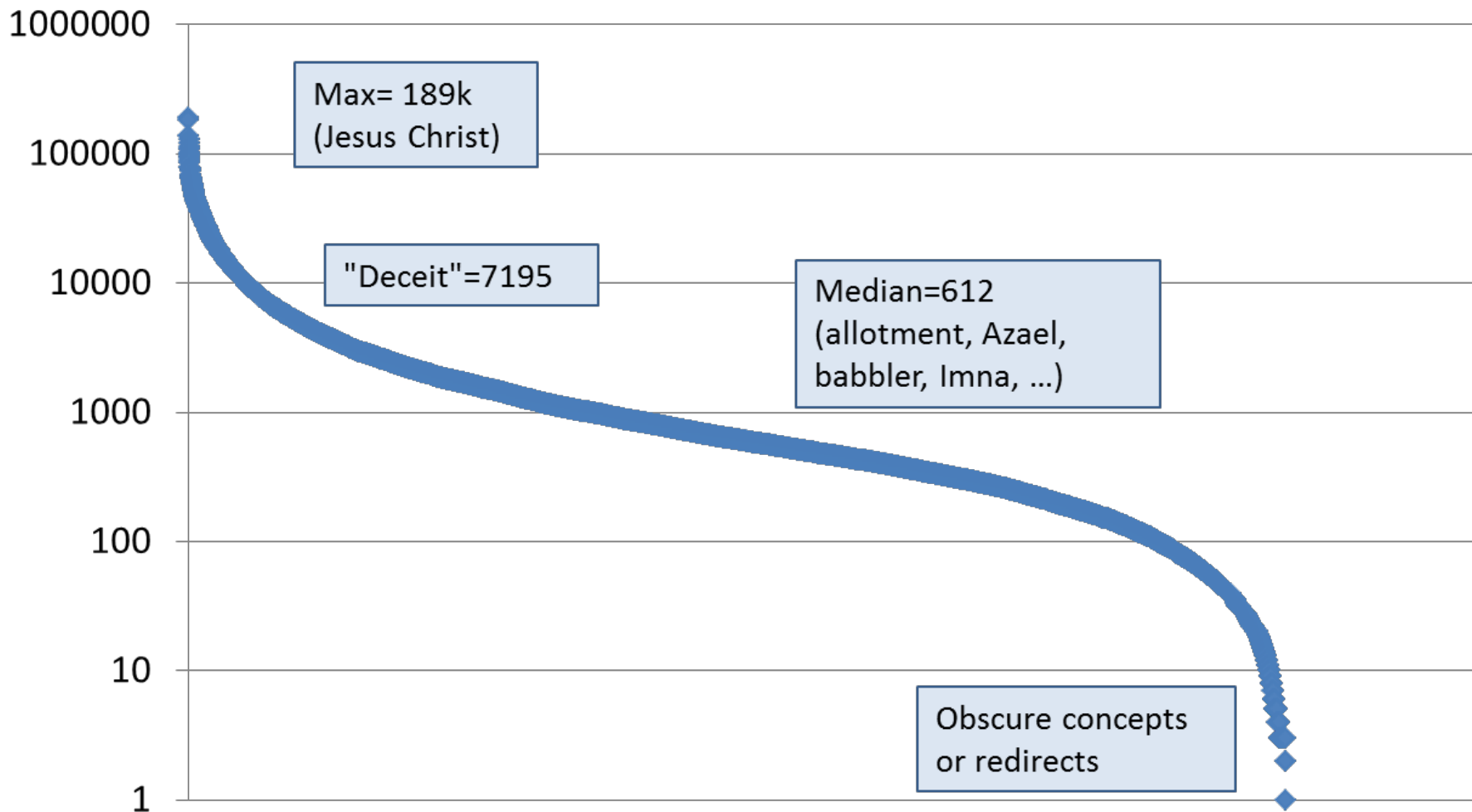
Zeros are either rare people or cases where we need to merge. There are no "true"

zeros.

The large majority have < 10 sources

So one definition of "important" is "broad"

Distribution of Terms for Concepts



Another definition of "important" is how much is written about it

I didn't freehand this! Gives some confidence that scale is helping here.

Other measures we don't yet have data for: popular (search, or topic guide), "significant" according to some other standard. But breadth and term count are

good cross-book approximations.

Learning Concept → Reference Associations

- Content about Biblical subjects typically includes references

Varies widely depending on the concept as well as the book

- Estimate the association between a concept and references by aggregating counts across all content (voting scheme)

This also allows us to rank them. Since we normalize references in our markup, they're language independent, so Spanish books help.

- These estimates should get better as

PRISCA [prɪsˈkə] (**Gk.** *Priska*), **PRISCILLA** [prɪ sɪlˈə] (*Priskilla*). A "tentmaker" (possibly leatherworker) known both as Prisca (**Rom. 16:3**; **1 Cor. 16:19**; **2 Tim. 4:19**) and Priscilla (a diminutive form, used in Acts), the wife of Aquila, also a "tentmaker." It has been suggested that Prisca was connected with the *gens Prisca*, a noble Roman family. She and Aquila left Rome because of Claudius' expulsion of the Jews from Rome and went to Corinth, where they hosted and worked with Paul (**Acts 18:2–3**). From Corinth they went with Paul to Ephesus where they explained "the way of God more accurately" to Apollos (**vv. 18–26**). After the death of Claudius they went back to Rome, where she and Aquila were the first to receive Paul's greetings from Corinth (**Rom. 16:3**).

Ac 18:26	20
2 Ti 4:19	19
Ro 16:3	16
1 Co 16:19	15
Ac 18:18	9
Ac 18:2	8
Ac 18:1-3	3

more content is added

Ac 18:2-3	3
Ro 16:3-5	3
Ro 16:4	3

We don't know why a reference is included, but idiosyncrasies tend to wash out with more data. In this example, there are only a few "canonical" references, so the distribution is sharp. That's less true for "fuzzier" concepts.

42 references, 137 instances

Learning Reference → Concept Associations

- Reversing this data gives evidence for concepts that are associated with a passage
- Each concept "votes" on the most likely concepts for a reference

Overlapping ranges are an equally thorny problem here.

tf.idf issues here: not raw frequency, but distinctiveness of the association between a concept and a reference

New Slide

COMMENT

- Bullet

Terms and References

COMMENT

- "Nebai" and "**Neh 10:19**" are both fairly unique content indicators
- Determining identity can be tricky:
 - Same word? (stemming or morphological analysis)
 - Same reference? (overlapping ranges)
- Word repetition is important evidence for topical relevance: reference repetition is rare.

Concept Relationships from Terms

Graphic of the Priscilla article from EERDMANS showing which frequent terms match other concepts. Green for a single preferred match, yellow for alt and/or ambiguous, red for no match?

- Bullet

New Slide

COMMENT

- Bullet

Reference Aggregation for %Prisca Content

Reference	Count	Reference	Count
Ac 18:26	10	2 Ti 4:19	7
I Co 16:19	4	Ac 18:18	4
Ro 16:3	3	Ac 18:2	3
Ro 16:3-5	3	Ro 16:4	2
Ac 18:25	2	Ac 18:2-3	2
I Co 1:12	2	I Co 16:12	2
Ac 18:11	2	Lk 8:3	1

Reference Aggregation for %Gabriel

Reference	Count	Reference	Count	Reference	Count
Lk 1:11-20	6	Tob 12:15	2	1 Th 4:16	1
Re 8:2	5	Ge 37:15	2	Da 10:5	1
Da 8:16	5	2 Ch 32:21	2	Lk 1:11	1
Da 9:21	4	Heb 7:25	1	Re 9:1-11	1
Lk 1:26-38	4	Da 10:21	1	Ex 3:2	1
Da 12:1	3	Lk 20:7	1	Da 8:18	1
Lk 1:19	3	Lk 1:31	1	Ro 14:10	1
Da 9:22	2	Da 9:21-22	1	Da 10:18	1
Lk 1:26-33	2	Da 9:21-27	1	Da 10:13	1
Dt 34:6	2	Re 8:6	1	Da 8:15-26	1

Challenges with Reference Counts

- How to combine counts
 - across overlapping references
 - for reference intersection with passage queries
- References might count less
 - if an article is longer
 - if the reference occurs later in an article
- Longer references are often more weakly "about" a concept
 - e.g. Is 40-66 for %Isaiah_Person
- We only know *that* a reference occurs, not why, or what it means
- We also manually provide "key references" for a small number of concepts

Reference-Based Retrieval: Acts 18.26

reference-based retrieval example

Reference-Based Retrieval: Luke 1.11-13

reference-based retrieval example

Concept Relationships through Reference Vectors

Do concepts of a feather reference together?

- Compare reference vectors to find concepts with similar references
- Example: ointment
 - Top reference: Mk 14:3 (n=19)
 - Top ten ointment references have 133 instances out of 672 (20%)
 - 92 other concepts have Mk 14:3 as a reference
 - Score them against the top ten ointment references

miracle	0.1977
GospelOfMark_Writing	0.1976
clean	0.1976
unclean	0.1976
joy	0.1975
Malachi_Person	0.1974
LordsSupper	0.1974
Peter_Person	0.1973
Gospel	0.1973
metal	0.1970

Top Ointment References

Mk 14:3	19
Lk 23:56	16
Es 2:12	14
Is 1:6	13
Ps 133:2	13
Ex 30:23-25	12
Job 41:31	12
Lk 7:46	12
Ec 7:1	11
Je 8:22	11

Conclusions

- You can learn

Resources and References

- Boisen, S (2007). [Deploying Semantic Technologies for Digital Publishing: A Case Study from Logos Bible Software](#). Presented May 24, 2007 at the Semantic Technology Conference.
- Boisen, S (2010). [A Controlled Vocabulary for Biblical Studies](#) . Presented March 27, 2010 at BibleTech:2010.

About This Presentation

- This is an XHTML document, structured and presented with [HTML Slidy](#). You can use your browser to view source and see how it works.
- Hyperlinks courtesy of [Logos Bible Software's RefTagger](#)
- The Logos editorial staff who have contributed to the Logos Controlled Vocabulary include Elizabeth Borries and Johnny Cisneros.