Video available on YouTube (with subtitles): https://www.youtube.com/watch?v=jEg1wPohXCo

UNIVERSITÄT OSNABRÜCK

Elen Le Foll

Issues in Compiling and Exploiting Textbook Corpora

October 2020

Japanese Association for English Corpus Studies

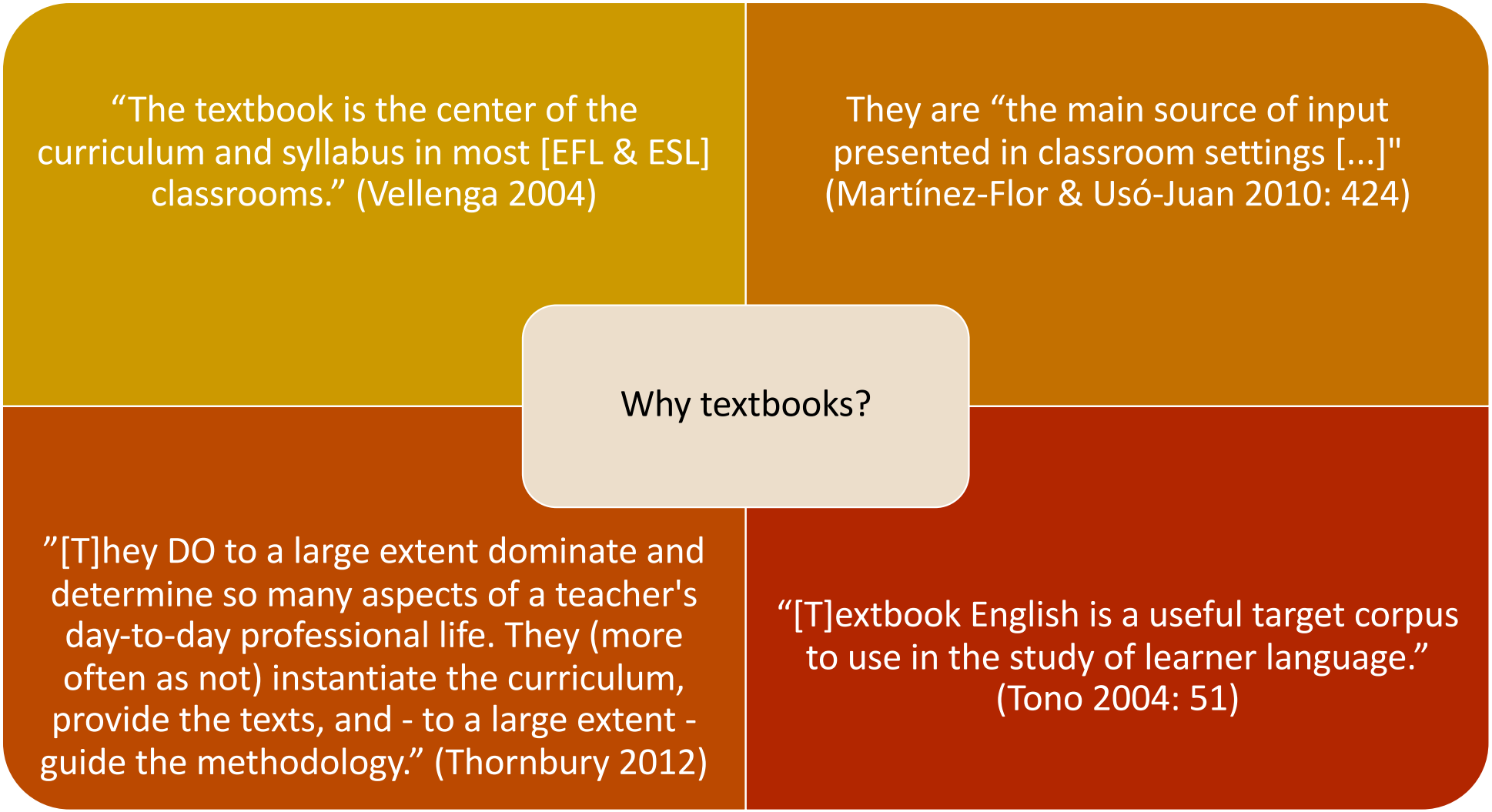
#JAECS2020

Video available on YouTube (with subtitles): https://www.youtube.com/watch?v=jEg1wPohXCo

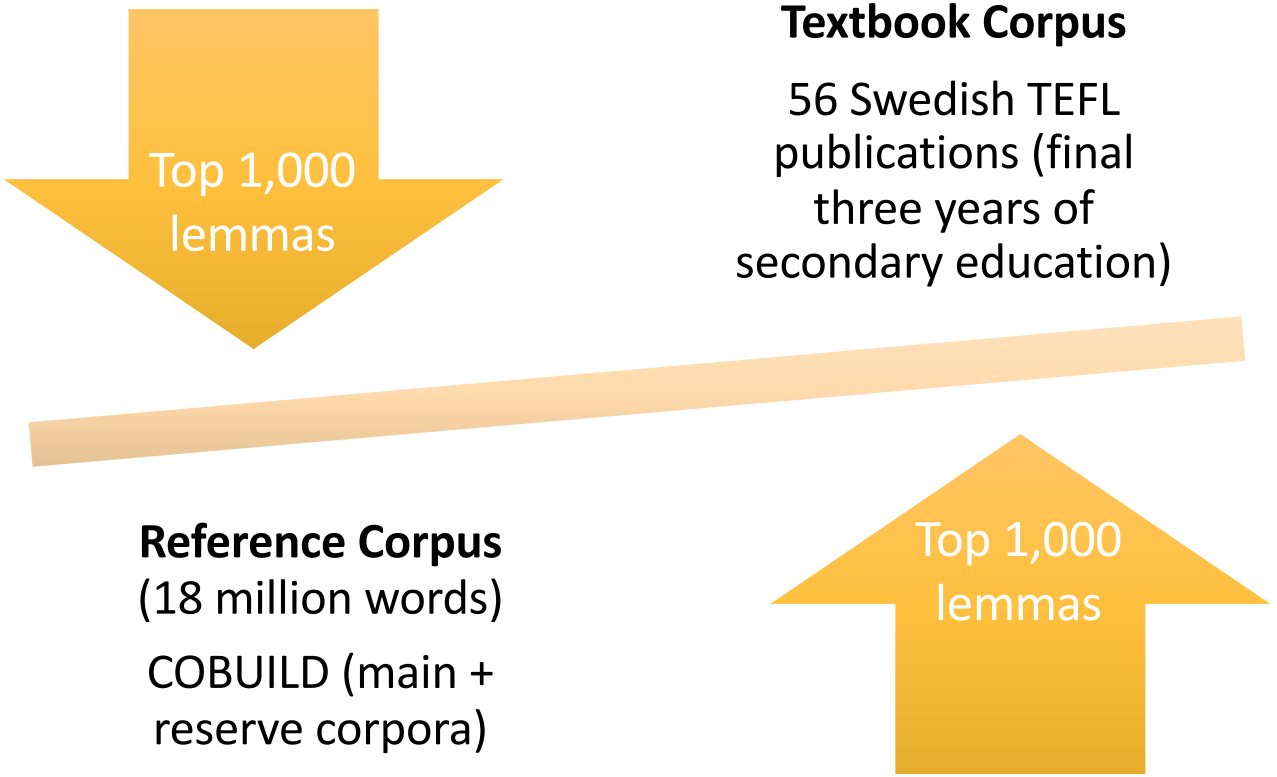
Issues in Compiling and Exploiting Textbook Corpora

- Potential of corpus-based textbook language studies
- Compiling a textbook corpus
  - Sampling frame
  - Selection process
  - Digitalisation and OCR
  - Mark-up and annotation
- Exploiting a textbook corpus
  - Register
  - Reference corpora
  - Text length
- Concluding discussion

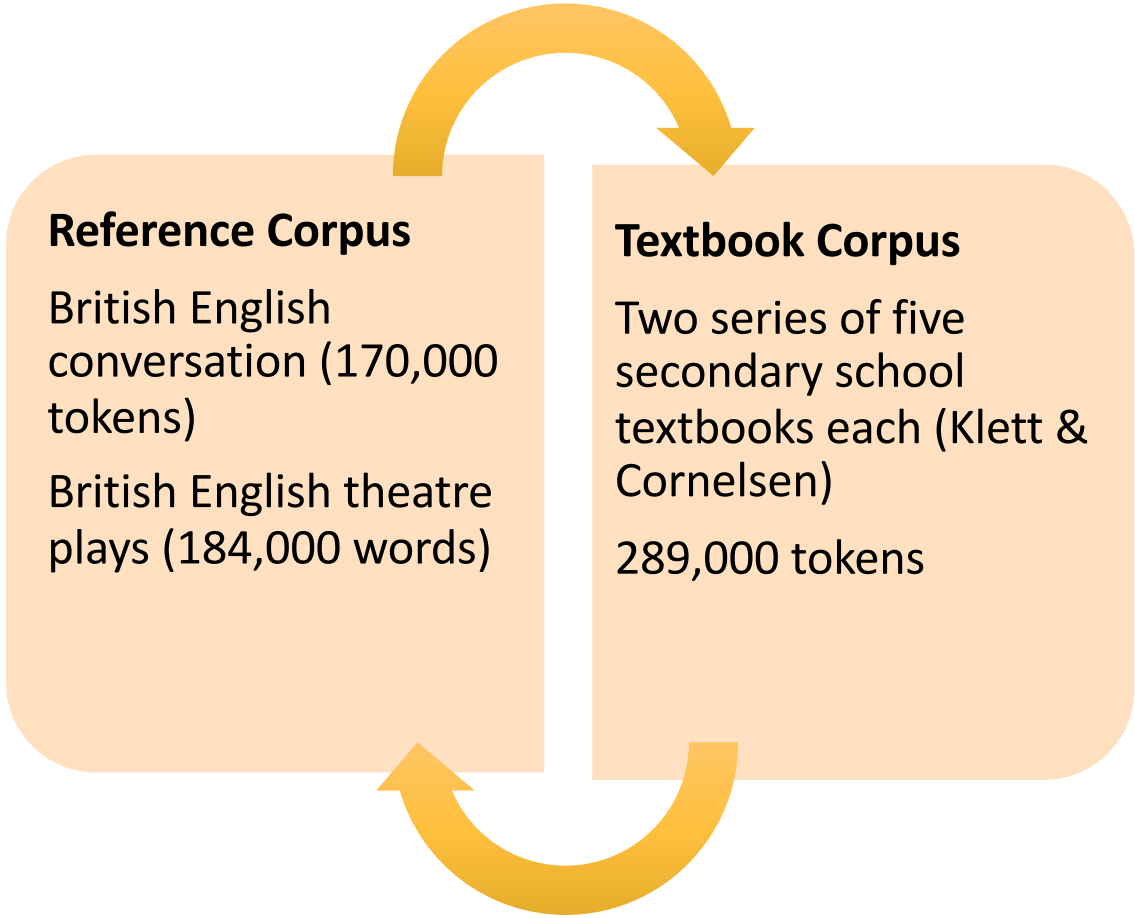
Slides download QR



Early Textbook Corpus studies: *Magnus Ljung (1990, 1991)*

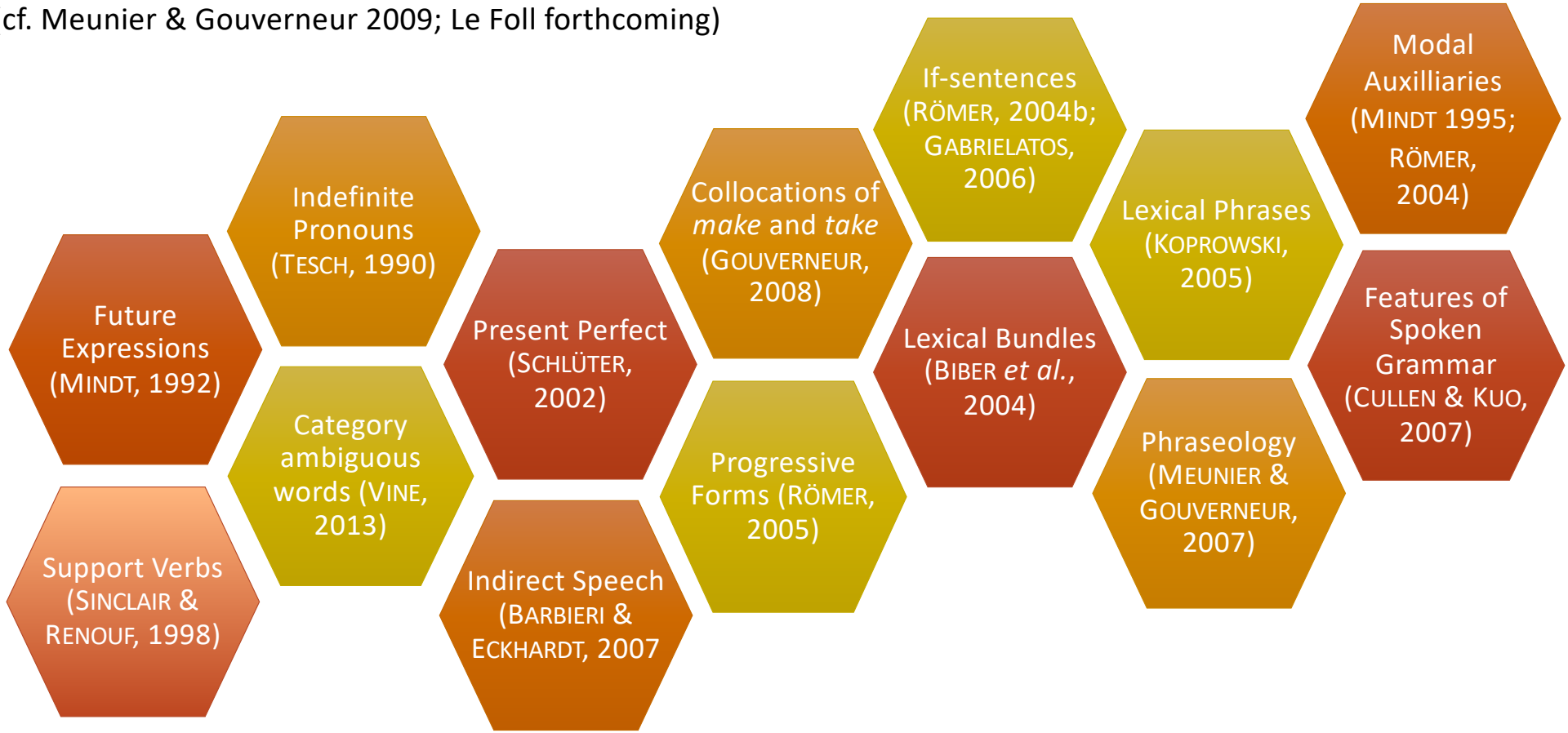


Early Textbook Corpus studies: *Dieter Mindt (1987, 1992)*



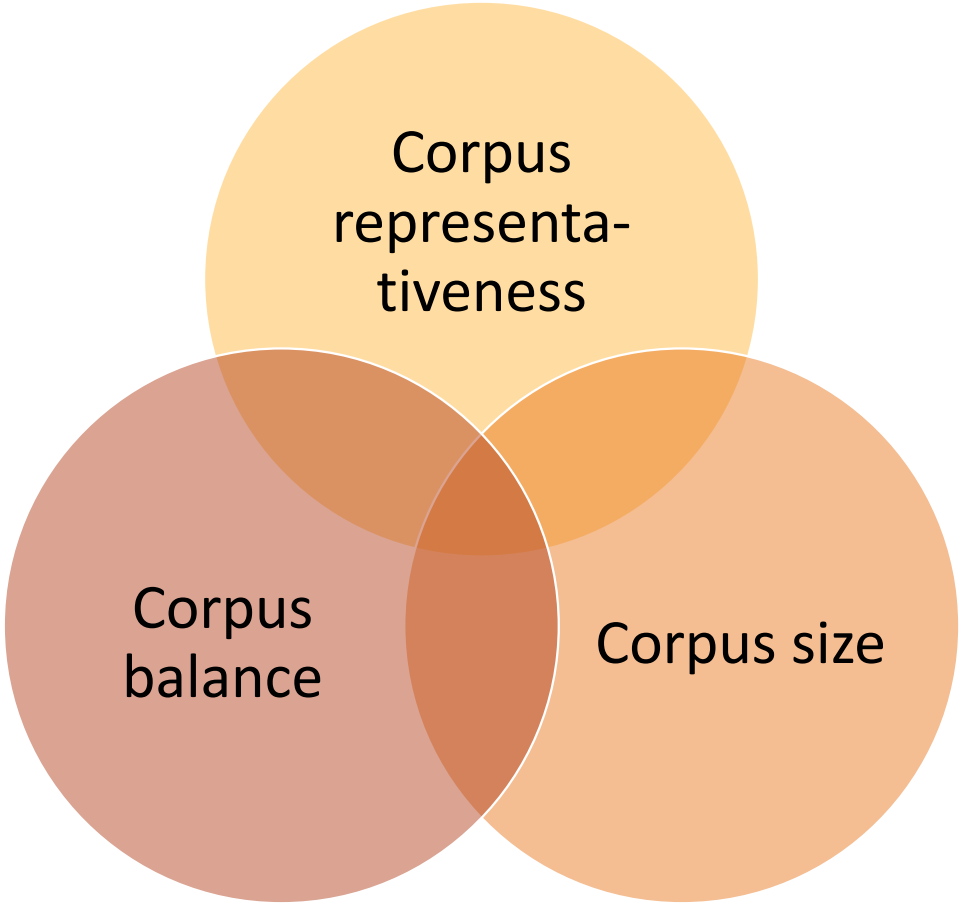
Corpus-based Textbook English studies

(cf. Meunier & Gouverneur 2009; Le Foll forthcoming)



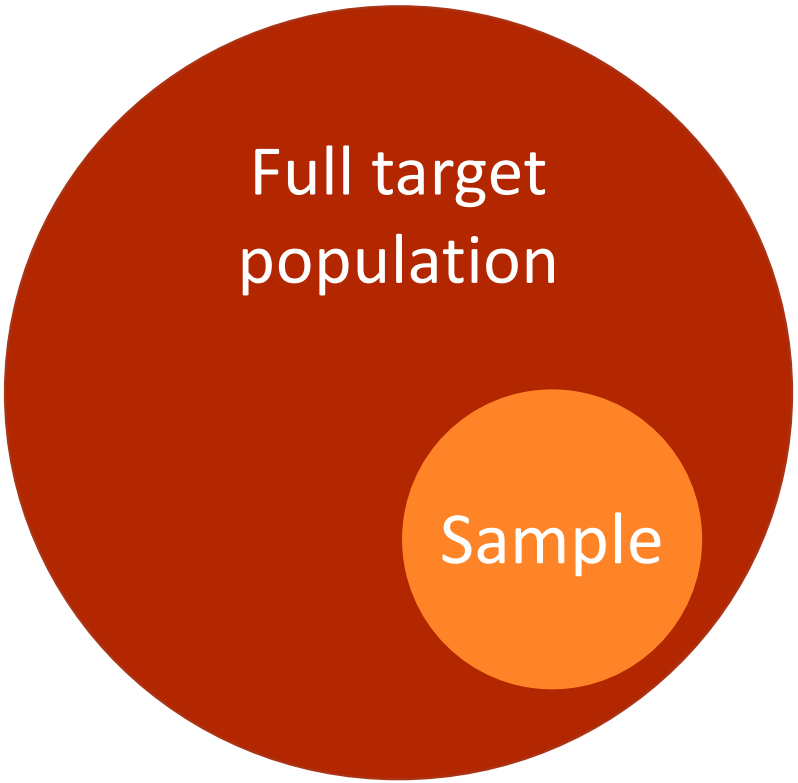
Selecting Textbooks

- Design of sampling frame



(cf. Biber 1993)

Selecting textbooks: Research question(s)



Target population:

English language content of **all** the **textbooks** from which all **lower secondary school students** in **France**, **Germany** and **Spain** were learning English as a **second or foreign language** between **2006 and 2018**.

Selecting textbooks: Definition

Green Line

Green Line – Bundesausgabe ab 2014

- Einstieg
- Konzeption
- Produktübersicht
  - Alle Lernjahre
  - 1. Lernjahr
  - 2. Lernjahr
  - 3. Lernjahr
  - 4. Lernjahr
  - 5. Lernjahr
- Lehrwerk-Online
- Testen und Fördern
- Stoffverteilung

Bundesland auswählen

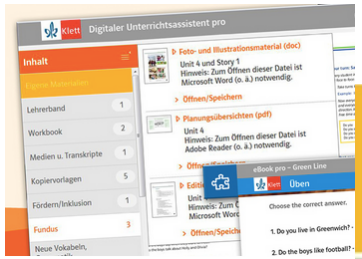
Schulart auswählen

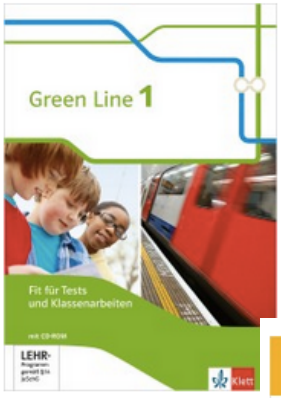
Fach auswählen

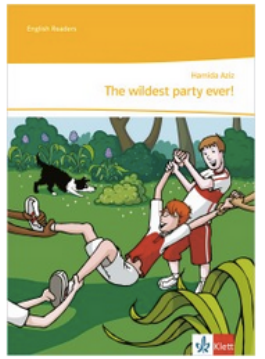
Produktübersicht

Schülerinnen und Schüler | Lehrerinnen und Lehrer

- **Lehrerbände (18)**
- **Digitaler Unterrichtsassistent (15)**
- **eCourse (1)**
- **eBooks (5)**
- **Kopiervorlagen und Arbeitsblätter (7)**
- **Lösungen (6)**
- **Leistungsmessung, Fördern, Differenzieren (22)**
- **Audiomaterialien (10)**
- **Videomaterialien (5)**
- **Software (15)**
- **Folien (4)**
- **weitere Materialien (2)**

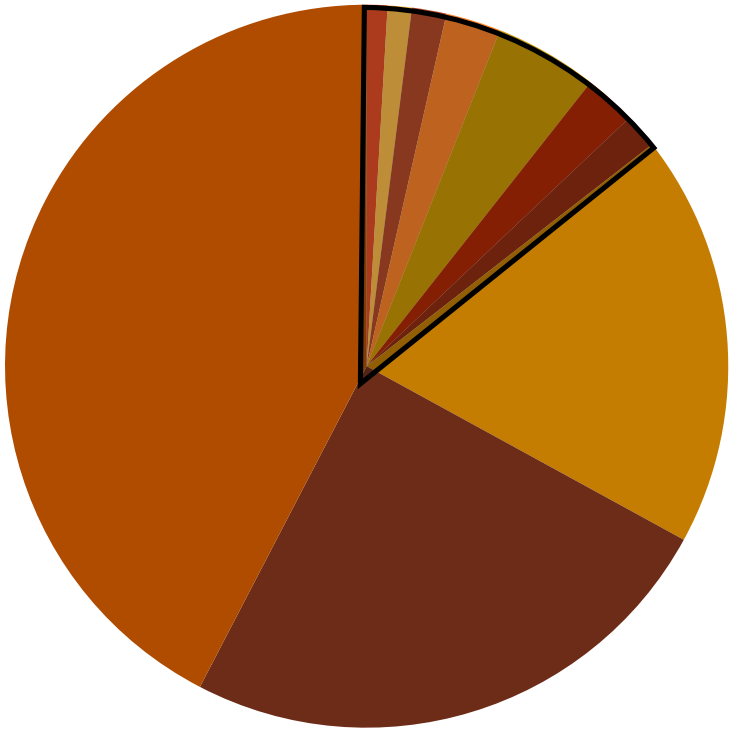






Green Line Vokabeltra...
Bildung

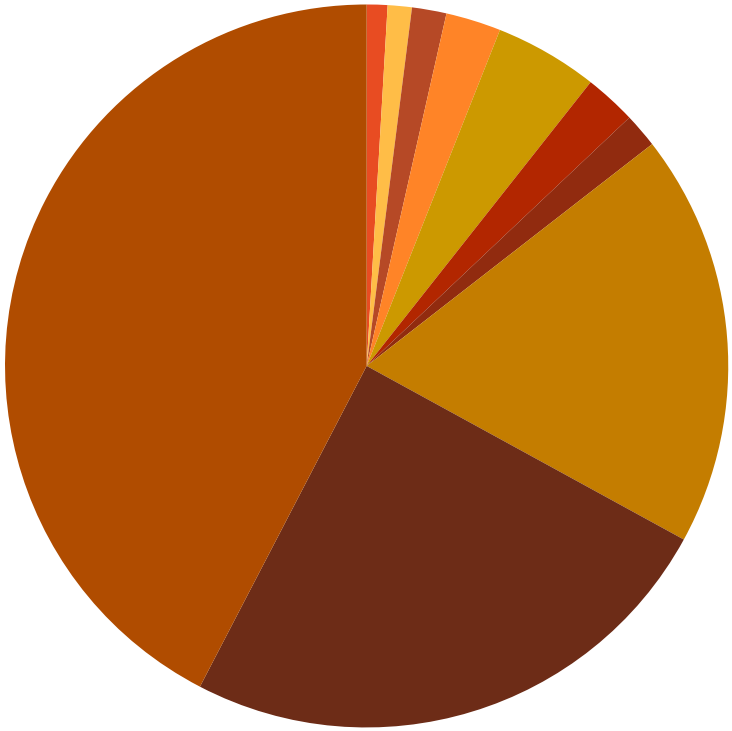
Selecting textbooks: Size and representativeness



Secondary school students learning English

- ➜ Textbook imposed by educational authority
- ➜ Publisher sale figures
- ➜ Informal survey of publishers
- ➜ Teachers/students
- ➜ Bookshops

Selecting textbooks: Balance



Secondary school
students learning English

➜ Different publishers

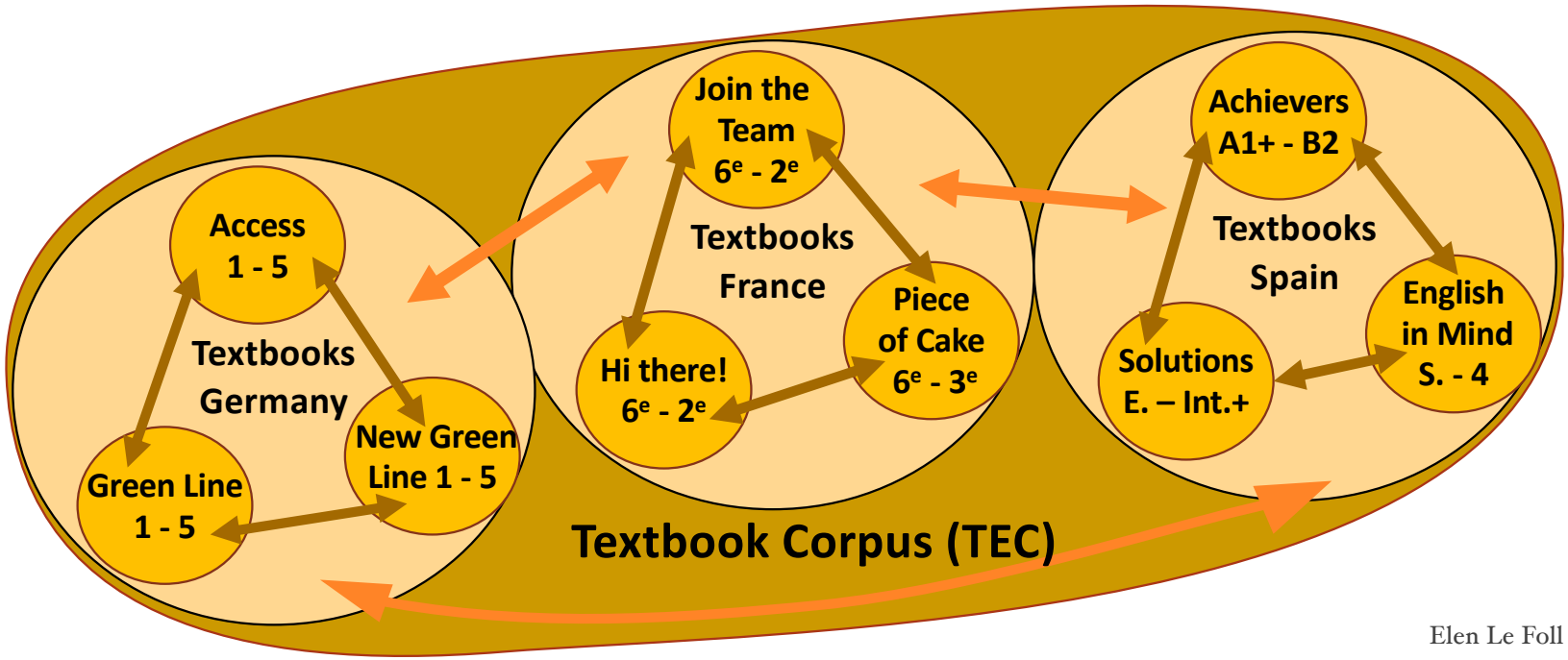
➜ Formats?

➜ Pedagogical approaches?

Selecting textbooks: Opportunist criteria

1. Availability of the textbooks in
   - Text/PDF format
   - Other digital formats (flash)
   - Print
2. Human resources
3. Financial resources

Textbook English Corpus (TEC; Le Foll forthcoming)



Digitalisation: OCR

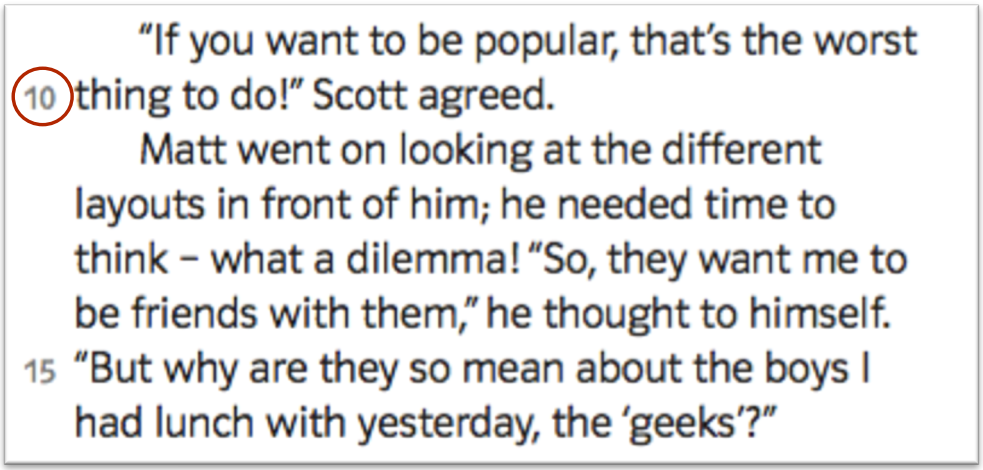
- Conversion to text files
- Complex OCR (optical character recognition) process:
  - Many different font types, colours
  - Complex page layouts
  - Multiple languages
- Automatic clean-up process:
  - Special characters, numbers, etc.
  - Regular expressions (python, R…)
  - SarAnt → www.laurenceanthony.net/software/sarant/

3 Keep calm and stop bullying

Activities

1. Read the survey and choose your answers anonymously.
2. Talk about the results of your class. See Grammar.
3. In groups, recap orally the results of your class using adverbs of frequency.
4. Group work Find solutions against bullying.
5. @Tricider Create a class survey with 5 new questions. See Grammar.

1. Has anyone ever insulted you?	Never	Once in a while	Too often
2. Has anyone ever hit you?	Never	Once	A few times
3. Has anyone ever threatened you?	Never	Seldom	Often
4. Has anybody been mean to you because of how you look?	Quite rarely	Once or twice	On a regular basis
5. Have you ever made mean comments on social networks?	Never	Sometimes	Too often
6. Have you ever received threatening text messages?	Once or twice	A few times	Too many times
7. Have you ever seen someone else being bullied?	Never	Often	Too many times
8. Have you ever threatened or pushed people? Or called them names?	No, I haven't	I'd rather not answer	Yes, I have

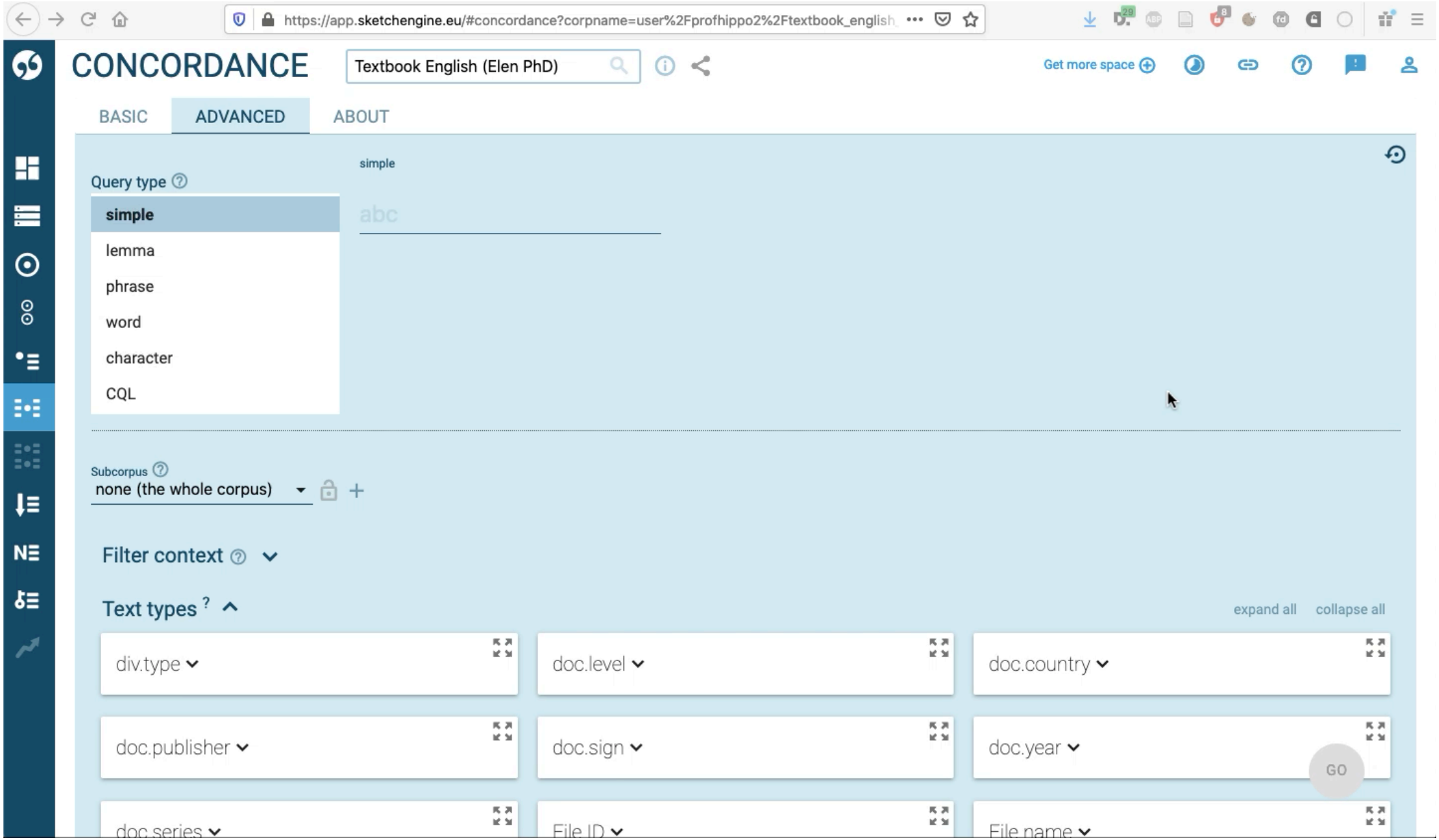


"If you want to be popular, that's the worst
10 thing to do!" Scott agreed.
Matt went on looking at the different layouts in front of him; he needed time to think – what a dilemma! "So, they want me to be friends with them," he thought to himself.
15 "But why are they so mean about the boys I had lunch with yesterday, the 'geeks'?"

Corpus mark-up: metadata

→ Aim for "a level of mark-up which maximizes the utility value of the text without incurring unacceptable penalties in the cost and time required to capture the data" (Atkins et al. 1992: 9, cf. Hardie 2014).

```
<doc sign="POC4" series="Piece of cake"
level="C" publisher="Livre scolaire"
year="2017" country="France">
```

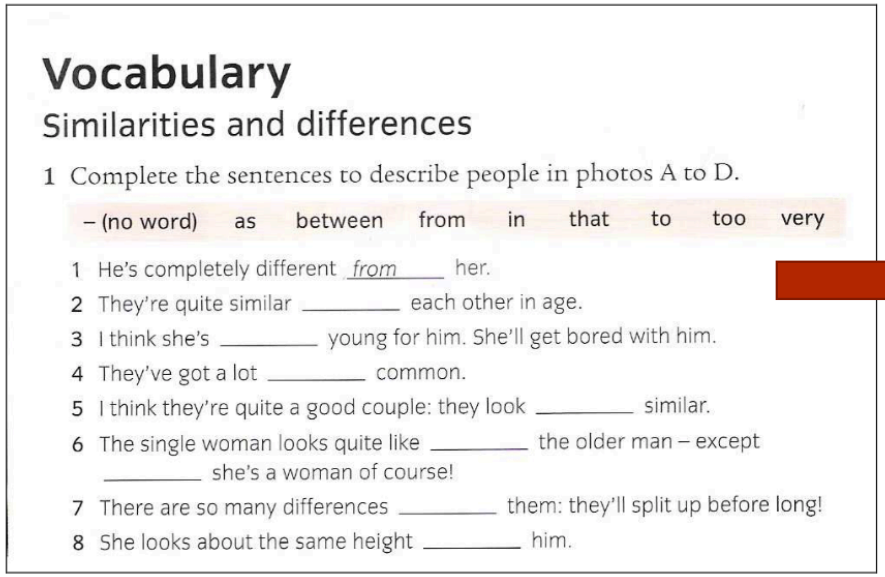

https://app.sketchengine.eu/#concordance?corpname=user%2Fprofhippo2%2Ftextbook_englis
CONCORDANCE
Textbook English (Elen PhD)
Get more space
BASIC
ADVANCED
ABOUT
Query type
simple
lemma
phrase
word
character
CQL
simple
abc
Subcorpus
none (the whole corpus)
Filter context
Text types
expand all
collapse all
div.type
doc.level
doc.country
doc.publisher
doc.sign
doc.year
GO
doc.series
File ID
File name

Automatic corpus annotation

- Part-of-Speech (POS) tagging
- Lemmatisation
- Syntactic dependency parsing

Manual corpus annotation

- TeMa Corpus (Meunier & Gouverneur 2009) → pedagogical tagging for vocabulary exercise subcorpus: 80+ tags for types of activities and status of lexical items



Vocabulary

Similarities and differences

1 Complete the sentences to describe people in photos A to D.

– (no word) as between from in that to too very

1 He's completely different *from* her.
2 They're quite similar _____ each other in age.
3 I think she's _____ young for him. She'll get bored with him.
4 They've got a lot _____ common.
5 I think they're quite a good couple: they look _____ similar.
6 The single woman looks quite like _____ the older man – except _____ she's a woman of course!
7 There are so many differences _____ them: they'll split up before long!
8 She looks about the same height _____ him.

Figure 2: Vocabulary exercise as it appears in the textbook

<CLISB-U6-P24-E1>

1213(BC)–#$
1213(BC)as#$
1213(BC)between#$
1213(BC)from#$
1213(BC)in#$
1213(BC)that #$
1213(BC)to#$
1213(BC)too #$
1213(BC)very#$

1213(CB)He's completely different 1213(AB)from# her$
1213(CB)They're quite similar 1213(AB)to# each other in age$
1213(CB)I think she's 1213(AB)too# young for him. She'll get bored with him$
1213(CB)They've got a lot 1213(AB)in# common$
1213(CB)I think they're quite a good couple : they look 1213(AB)very# similar$
1213(CB)the single woman looks quite like 1213(AB)no word# the older man-except 1213(AB)that# she's a woman of course !$
1213(CB)there are so many differences 1213(AB)between# them ; they'll split up before long !$
1213(CB)She looks about the same height 1213(AB)as# him$

Figure 3: Pedagogically annotated vocabulary exercise as it appears in the corpus

Manual corpus annotation

- TEC (Le Foll forthcoming) → text register

Textbook Registers

S1/47 L2/19

Well, what's your idea?

After school on Friday …

Dave: Hey you two. Let's do something fun together at the weekend.
Jay: Great idea, Dave! But Luke: No football, OK?
Luke: OK, OK. Do you like swimming?
Jay: Yes, I do. Swimming is fun.
Luke: We can go to Arches Leisure Centre.
Dave: Not so fast! I don't like swimming and water slides.
Luke: Oh, sorry. Well, what's your idea?
Dave: What about something special? Do you know the Cutty Sark museum?
Jay: No, I don't. But is a *museum* cool?
Dave: *Very* cool – and exciting too.
Luke: Great. So where can we meet? And what time?
Dave: Let's meet at my house on Sunday. I live on the way to Cutty Sark.
Jay: Is 11 o'clock OK? In my family we don't get up very early on Sundays.
Dave: Yes, 11 o'clock at my house is fine. OK, see you on Sunday morning!

WELCOME TO MUDCHUTE PARK & FARM

Mudchute Park and Farm are free for everyone. Our farm is one of the biggest inner-city farms in Europe. **Visitors** can see over 200 animals here: large farm animals in the **fields**, horses and ponies in the **Riding Centre** and small animals in Pets Corner. If you want to visit Pets Corner, have a **tasty** snack at Mudchute Kitchen or **book** a riding lesson, please check the **opening times** below.

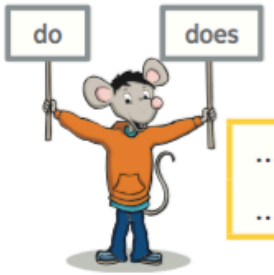
Fun in the Riding Centre

Mudchute Park & Farm: Farm: Tuesday–Sunday 9–5; Park: **All day** every day
Pets Corner: Monday–Sunday 9–4 **Mudchute Kitchen**: Tuesday–Sunday 9–5
Riding Centre: Monday–Thursday 8–9; Friday 8–4:30; Saturday–Sunday 8–5:30
How to find us: We are on the Isle of Dogs. It is easy to get here by car, train, bus (D3, D6 or 135 bus), on foot, by bike, **river ferry** or the Docklands Light Railway (the best DLR **station** is Crossharbour).

LANGUAGE 7

Do or *does*? → G17

a) *Complete the rule. Put the new rule with your rule from page 70, Ex. 2.*



…	I, you, we, they	verb	rest of sentence
…	he, she, it		

b) *Put in the correct forms.*

1. … the website give lots of information?
2. … it show photos with rabbits?
3. … Claire and Desmond ask questions?
4. … they think the idea of Mudchute is crazy?
5. … lots of animals live on the farm?
6. … Olivia like picnics?

Dear Rob,

I'm sorry, but I can't come at 1:30.
Please come to my pizza party on Saturday 17th March at 1:30 p.m.
Thank you for your invitation.
Love, Rob
My address is 16 Rosendale Road.
Dear Lara,
Is that OK?
Let me know if you can come.
Love, Lara
I would love to come a little later.
P.S. You must wear a silly hat!

This is my mum, Sally is her name,
This is my dad, football is his game,
This is my sister, she's a crazy kid.
Together we are a family.

Honey's in the kitchen, Mr Fluff is with her too,
They like our flat – what about you?
I love my pets, and they love me,
Together we are a family.

Chorus:
Mum and Dad, my sister and me,
Together we are a family.
Mum and Dad, my sister and me,
This is my family.

Hi, I'm Holly. What's your name?
Hello, nice to meet you.
Where are you from?
Oh, that's great!
Hey, can you sing? Yes?
Let's sing together!

(Chorus)

Musik und Text: T. Dorsch / P. Hoke

Does the farm look nice?

"Do you spell it with S?" Olivia asks.
"No, you don't. You spell it with C," her father Desmond tells her.
At last Olivia finds the website for Mudchute Farm.
"Does the farm look nice?" Claire asks.
"Yes, it does," Olivia says. "They've got lots of animals."
"Have they got rabbits?" Olivia's half-sister Lucy asks.
"Yes, there are rabbits at Pets Corner. Look, I can show you a photo of them."
Lucy is very happy. "I love rabbits and rabbits love me!" she says.
The website gives lots of information. The farm doesn't open on Mondays, but it's open all other days of the week. It's easy to get there from Greenwich too. They can go by DLR (Docklands Light Railway).
"But does it cost lots of money?"
"No Dad, it doesn't. It's free. And we can take our own picnic," Olivia says.
The Frasers like the idea of Mudchute Farm. "What about next Sunday?" Olivia asks. "Maybe Holly can come with us."
"Yes, ask her," Claire answers. "But Sunday is often a bad day. *Lots* of people."
"Claire is right. Let's go on Saturday," Desmond says. "But we want good weather – so keep your fingers crossed!"

die Regeln aufstellen.

- *Um ein* **Verbot** *auszudrücken, kannst du auch* **mustn't** *verwenden:*
  You **mustn't** talk to Holly like that. — *Du* ***darfst*** *mit Holly* ***nicht*** *so reden.*
- *Mit der* **Frage Can you ...?** *leitest du eine höfliche Bitte ein.*
  Du erkennst sie auch am Wort please. *Solche Fragen werden oft mit einer Kurzantwort beantwortet.*
  Holly: **Can you** make a cake, please? — ***Kannst du*** *bitte einen Kuchen backen?*
  Mum: **Yes, I can./No, I can't.** — ***Ja./Nein.***
- *Die Modalverben* can, can't *und* mustn't *sind* ***in allen Personen gleich.***
  Du verwendest sie nur für die Gegenwart.
- *Dem* **Modalverb** *folgt immer ein* **Vollverb** *in der Grundform (außer in der Kurzantwort):*
  Holly **can** **invite** her friends.

Mustn't *bedeutet also nicht „nicht müssen", sondern* ***„nicht dürfen".***

Manual corpus annotation

- TEC (Le Foll forthcoming) → text register

Low cost annotation solution: Keyboard Maestro

```
Try to guess what each piece of information refers to. Cmd + I
```

```
<div type="instructional"> Try to guess
what each piece of information refers
to. </div>
```

Preview File Edit View Go Tools Window Help 100% Thu 10:02

New_GreenLine_4_PDF.pdf (page 1 of 286)

Search

New_GreenLine_4_PDF.pdf

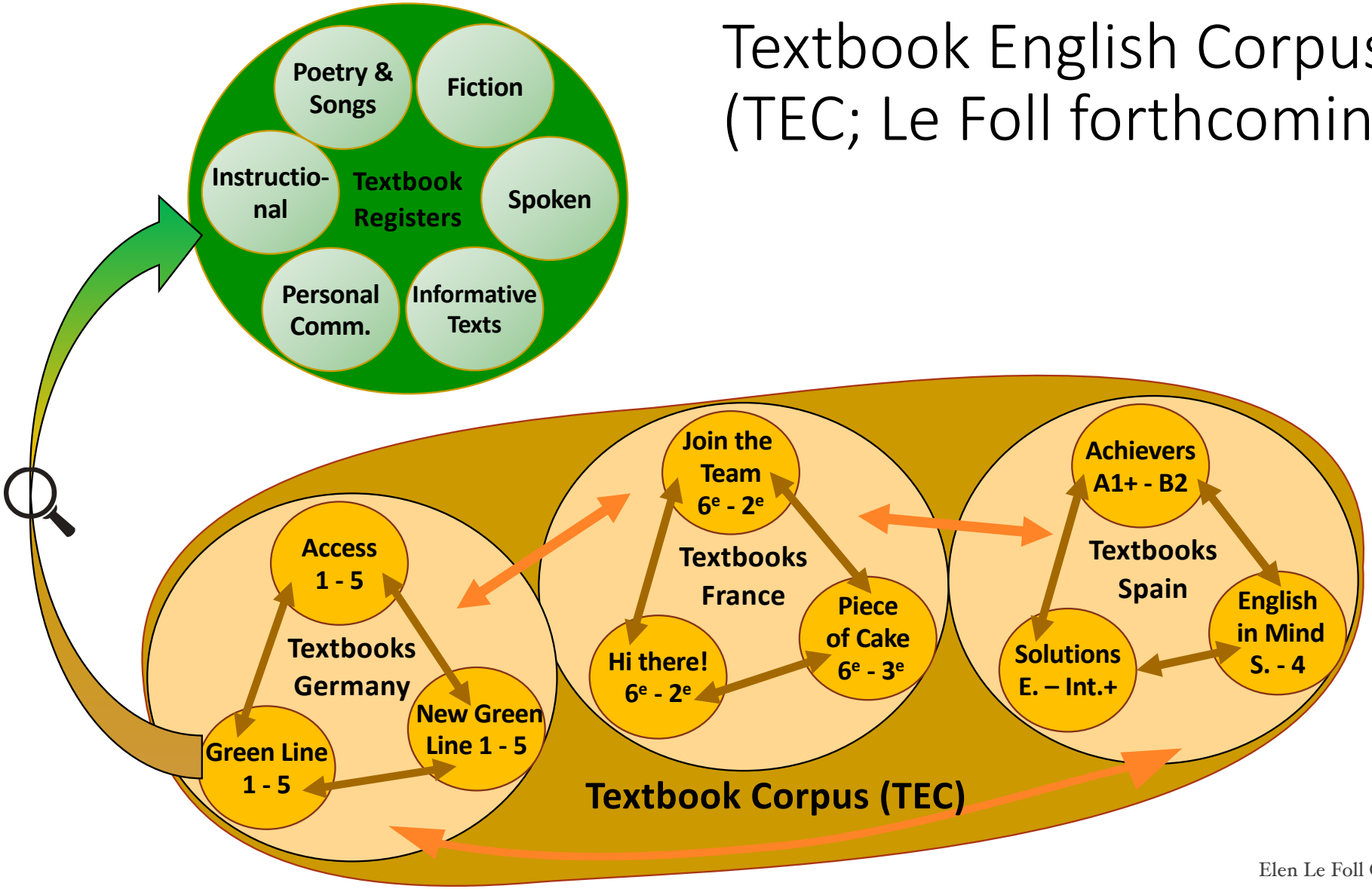
12

13

14

Green Line 4

Textbook English Corpus (TEC; Le Foll forthcoming)



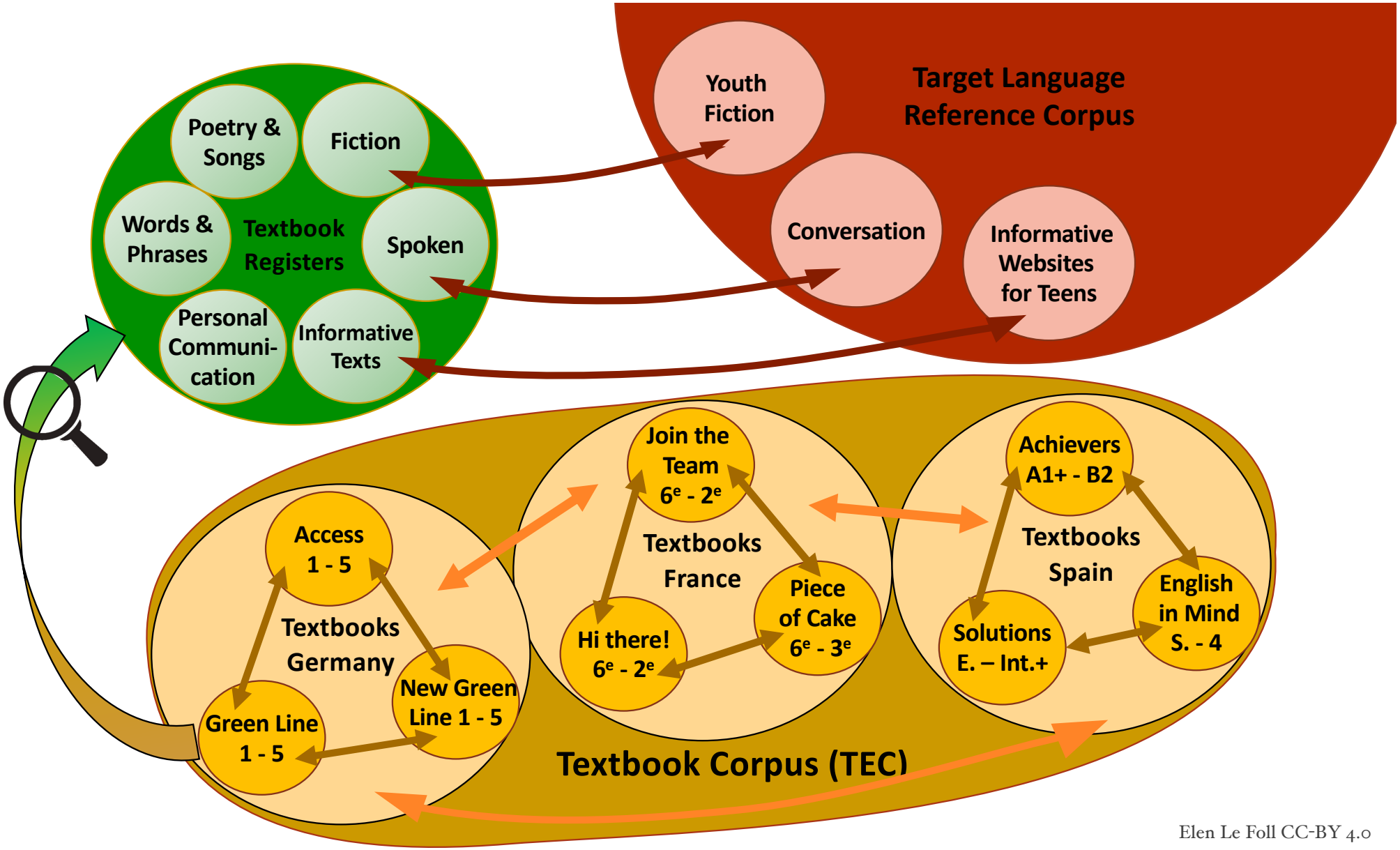
Target learner language

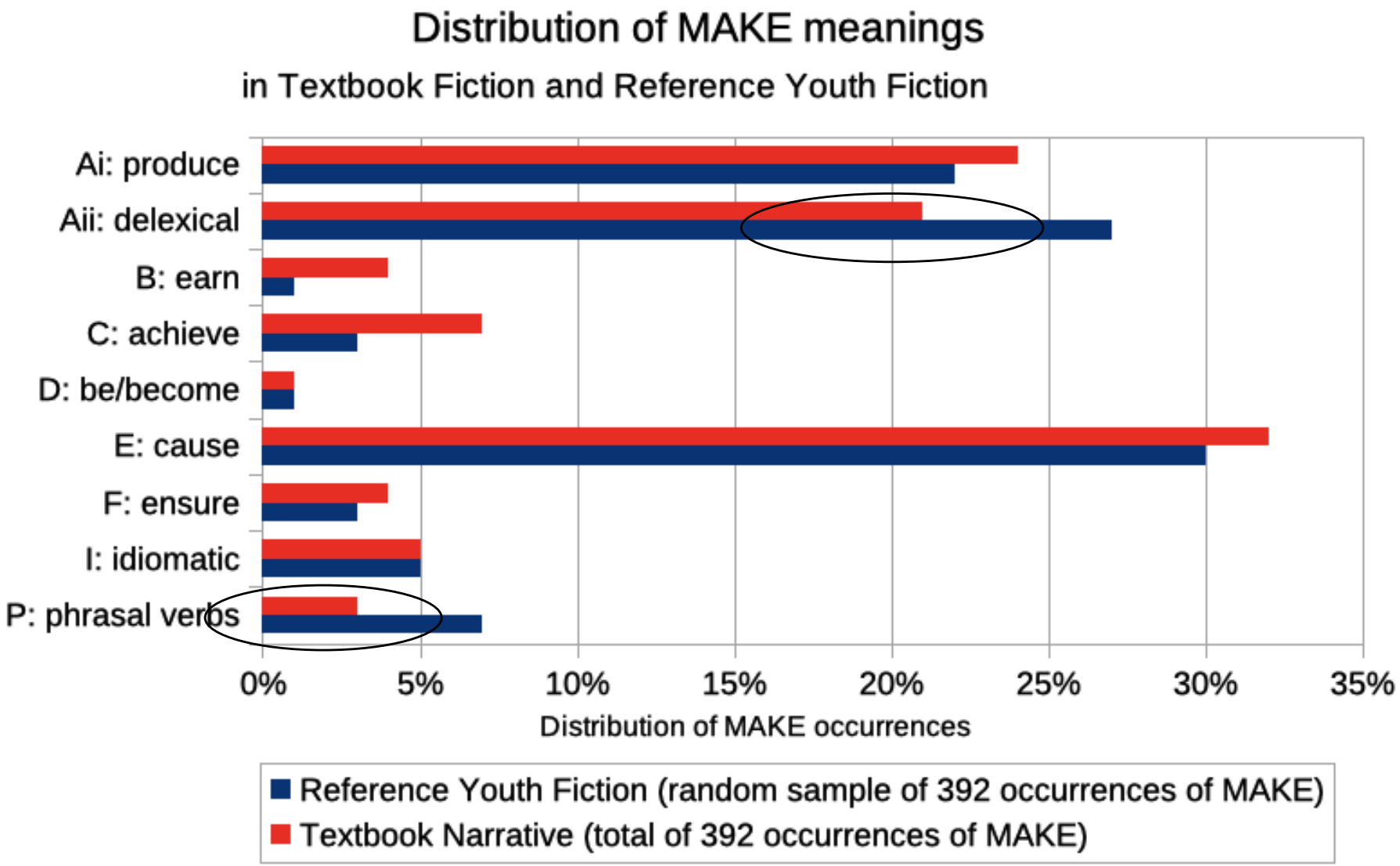
Target Language Reference Corpora

Conversation

Youth Fiction

Informative Websites for Teens





Collocates of delexical MAKEs

Youth Fiction Reference

Textbook Fiction

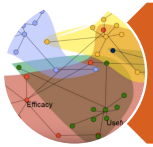
Multi-Dimensional Analysis (MDA)

Biber (1988): Variation across speech and writing

Compilation of large corpus representing full register varieties

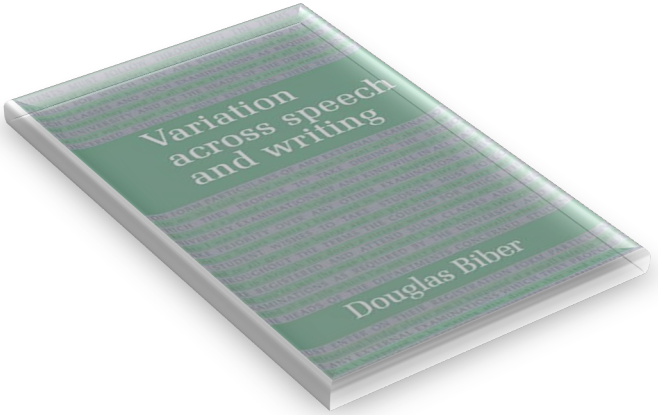
Tag and count broad range of linguistic features in all texts

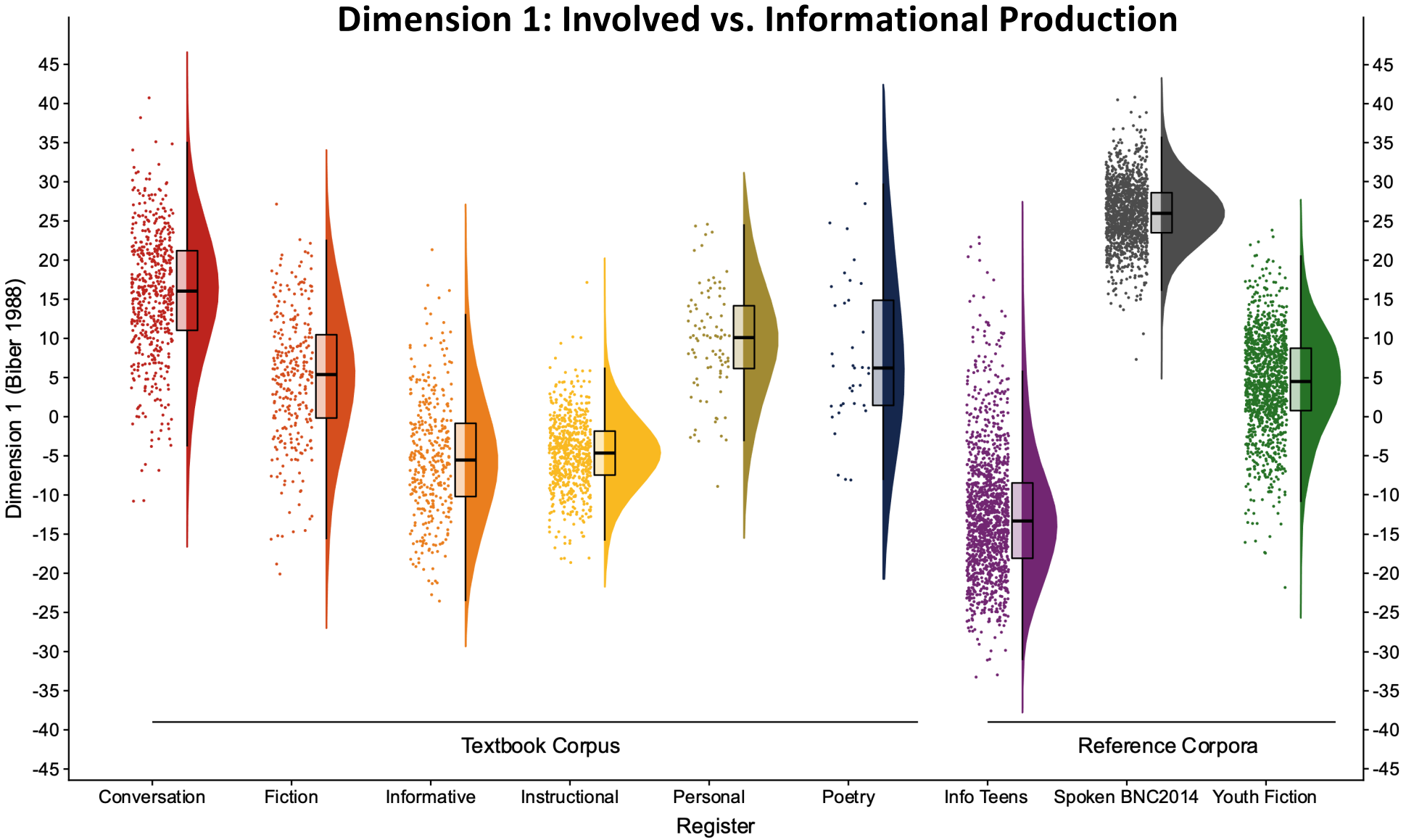
Correlation matrix of 67 features across all the texts

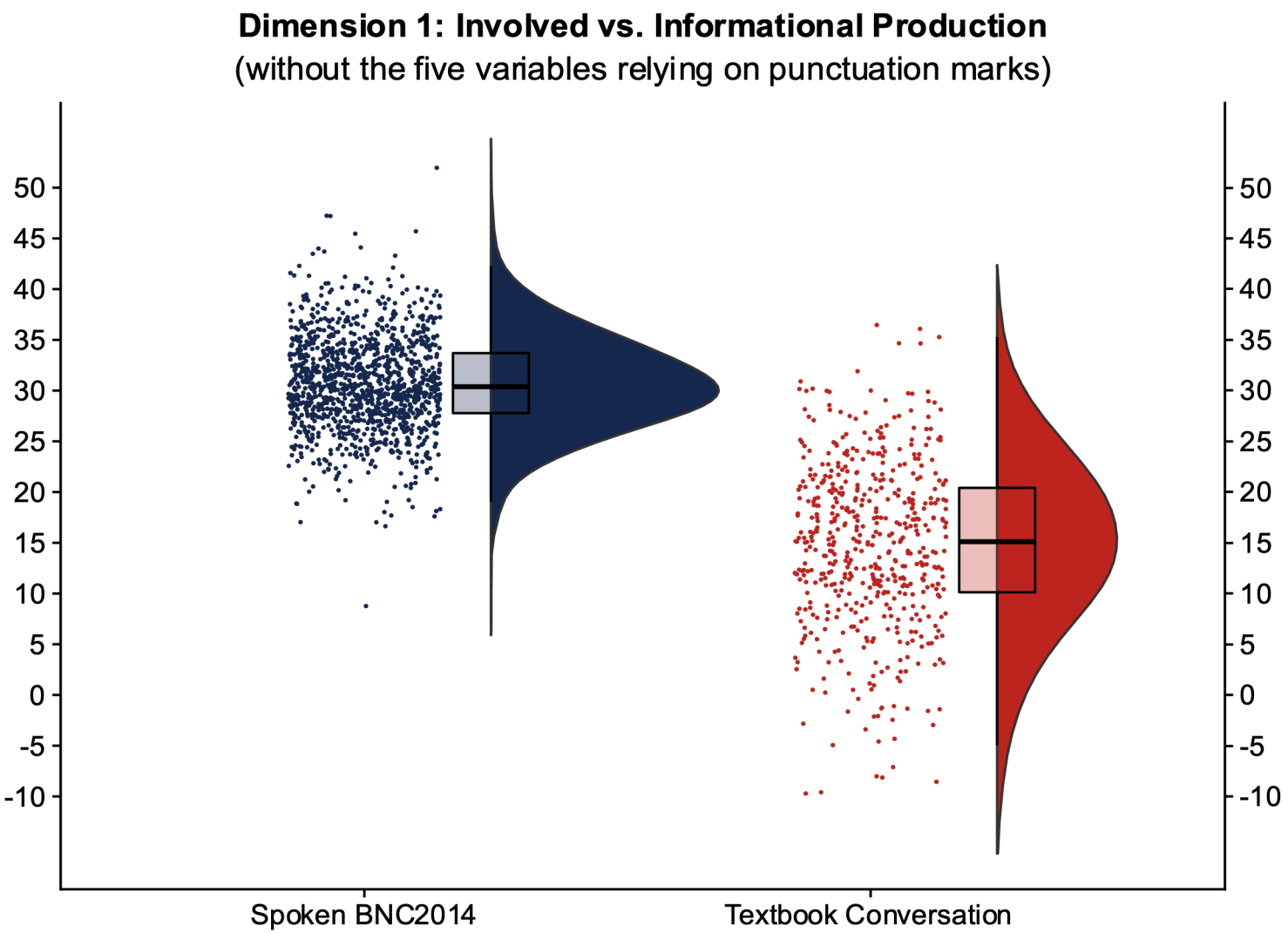


Factor analysis to extract max. amount of shared variance

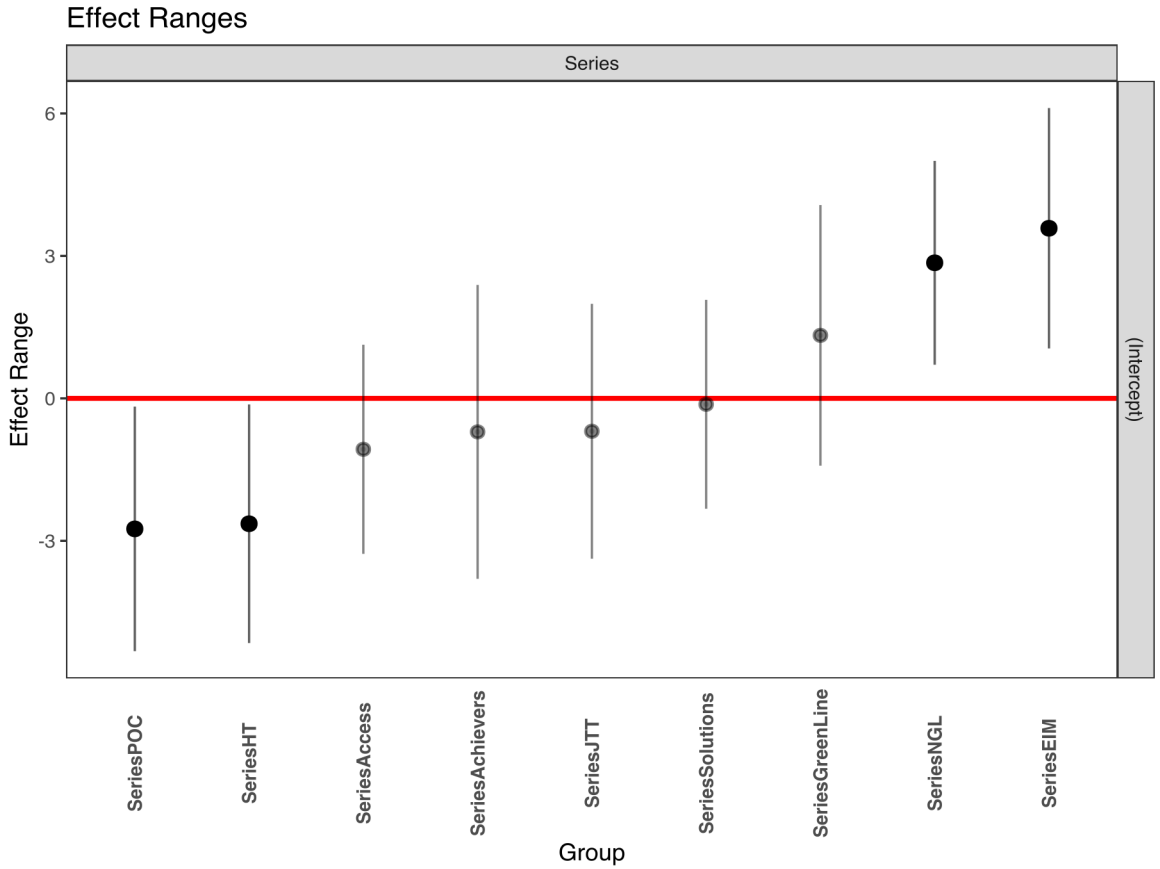
Functional interpretation of factors as six dimensions of register variation in English

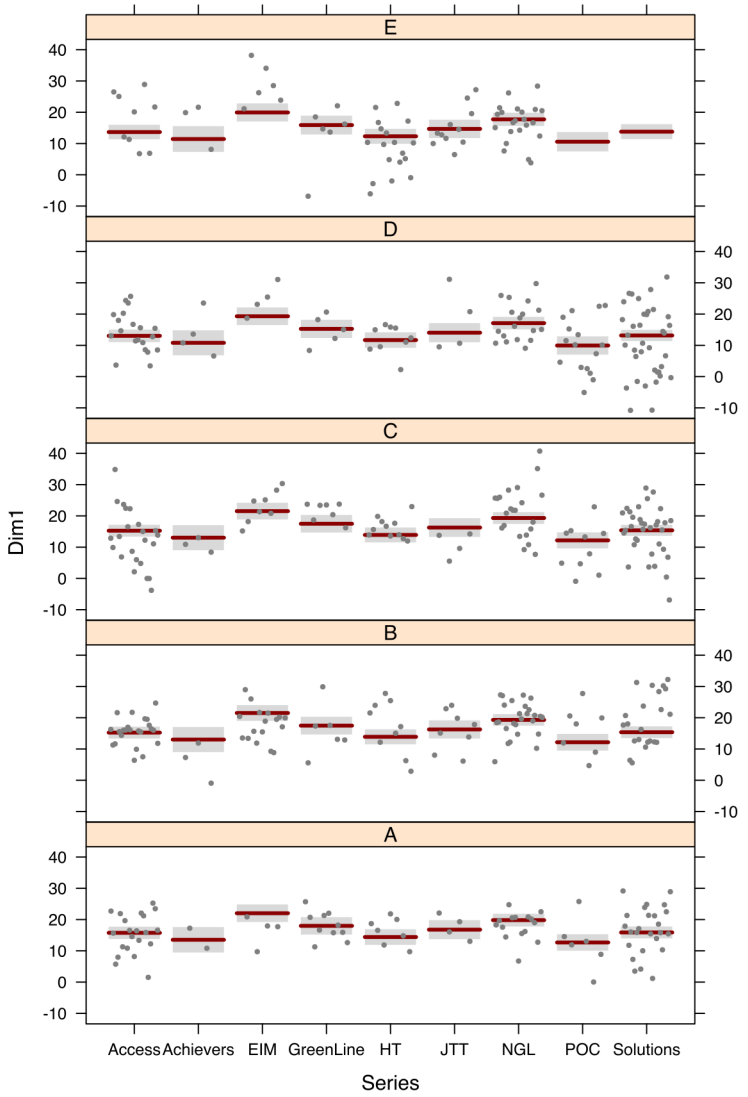



Dimension 1: Involved vs. Informational Production
Dimension 1 (Biber 1988)
45
40
35
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
Textbook Corpus
Reference Corpora
Conversation
Fiction
Informative
Instructional
Personal
Poetry
Info Teens
Spoken BNC2014
Youth Fiction
Register



Dimension 1: Involved vs. Informational Production
(without the five variables relying on punctuation marks)





Textbook conversation

Dad: The **phone** is ringing. I am sure it's **mum** calling **from Auckland**.

Jimmy: Hello, **mum**! How are you?

Mum: Oh I'm fine! What about you? Tell me **about** your **expedition**.

Jimmy: It's great! I love it! We've sailed more than 80 **miles** since Portsmouth... The **weather** has been great so far... I've seen **whales** and **dolphins**... **Mum**, they are so beautiful... I've fed **huge birds** and I've caught **huge fish**....

Mum: And do you help **George** and **Dad on** the **boat**?

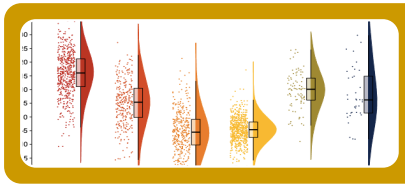
Jimmy: Oh sure! I am a **real sailor** now! I've just finished cleaning the **deck**.

<Join the Team 4e>

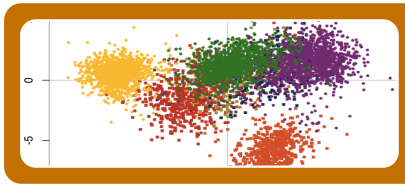
Naturally-occurring conversation

- yes **it's it's** erm something from the greenflies I **think** rather than it**'s not** the tree itself **it's** the fact that **it's** the aphids erm producing something
- do you **think [THATD]** they drink too drink too much of **this** and **it** makes them ill?
- I **think [THATD]** they go they go too too mad on the on the sap and **it just** produces all **this** sticky goo
- oh gosh I **didn't know** *<BNC2014 SRWD>*

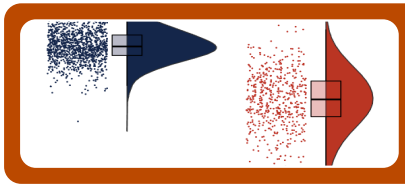
Preliminary findings from the TEC



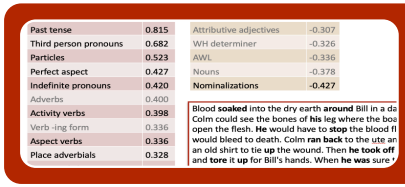
→ “Condensed” register variation in Textbook English



→ Instructions and explanations (1/3 word count of textbook texts) = very distinct register



→ Poor representation of spontaneous conversation



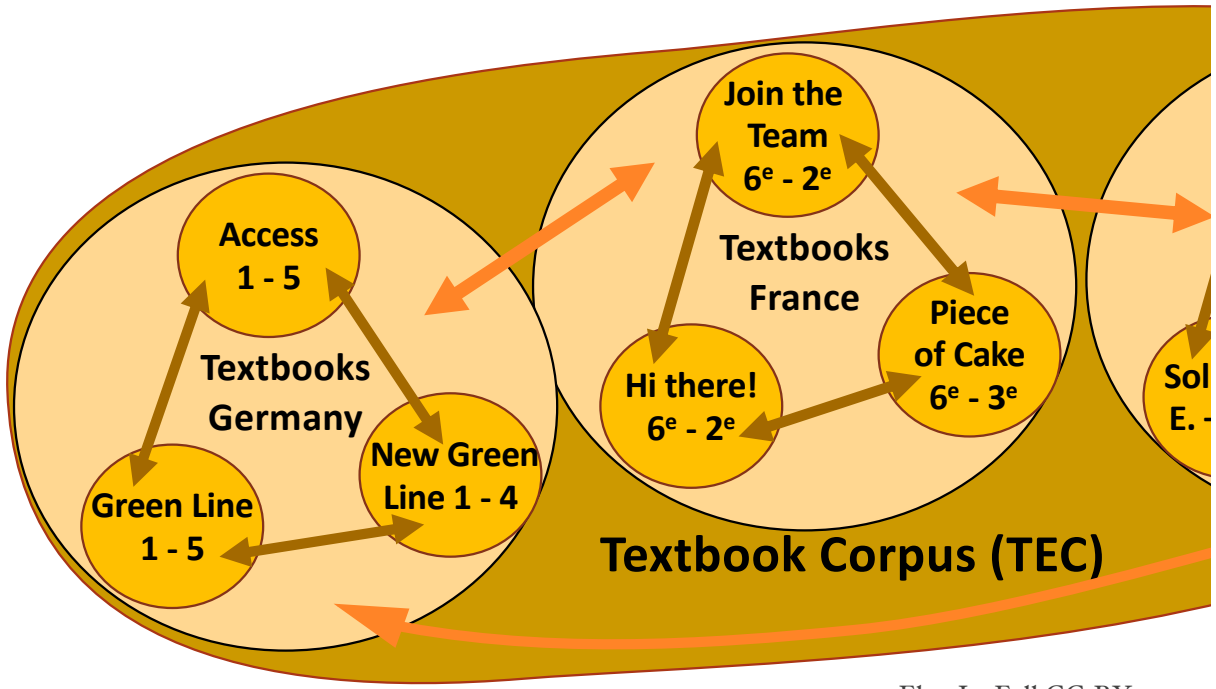
→ More accurate representation of fiction

(cf. Le Foll 2018; 2019; 2020; forthcoming)

www.youtube.com/watch?v=sePgPsspVE8

Issues in Compiling and Exploiting Textbook Corpora

- Potential of corpus-based textbook language studies
- Compiling a textbook corpus
  - Sampling frame
  - Selection process
  - Digitalisation and OCR
  - Mark-up and annotation
- Exploiting a textbook corpus
  - Register
  - Reference corpora
  - Text length
- Concluding discussion



Conclusion

- Text retrieval from flash files
- No direct link to actual textbooks (and formatting not encoded)
- Text length
- Open access → copyright issues

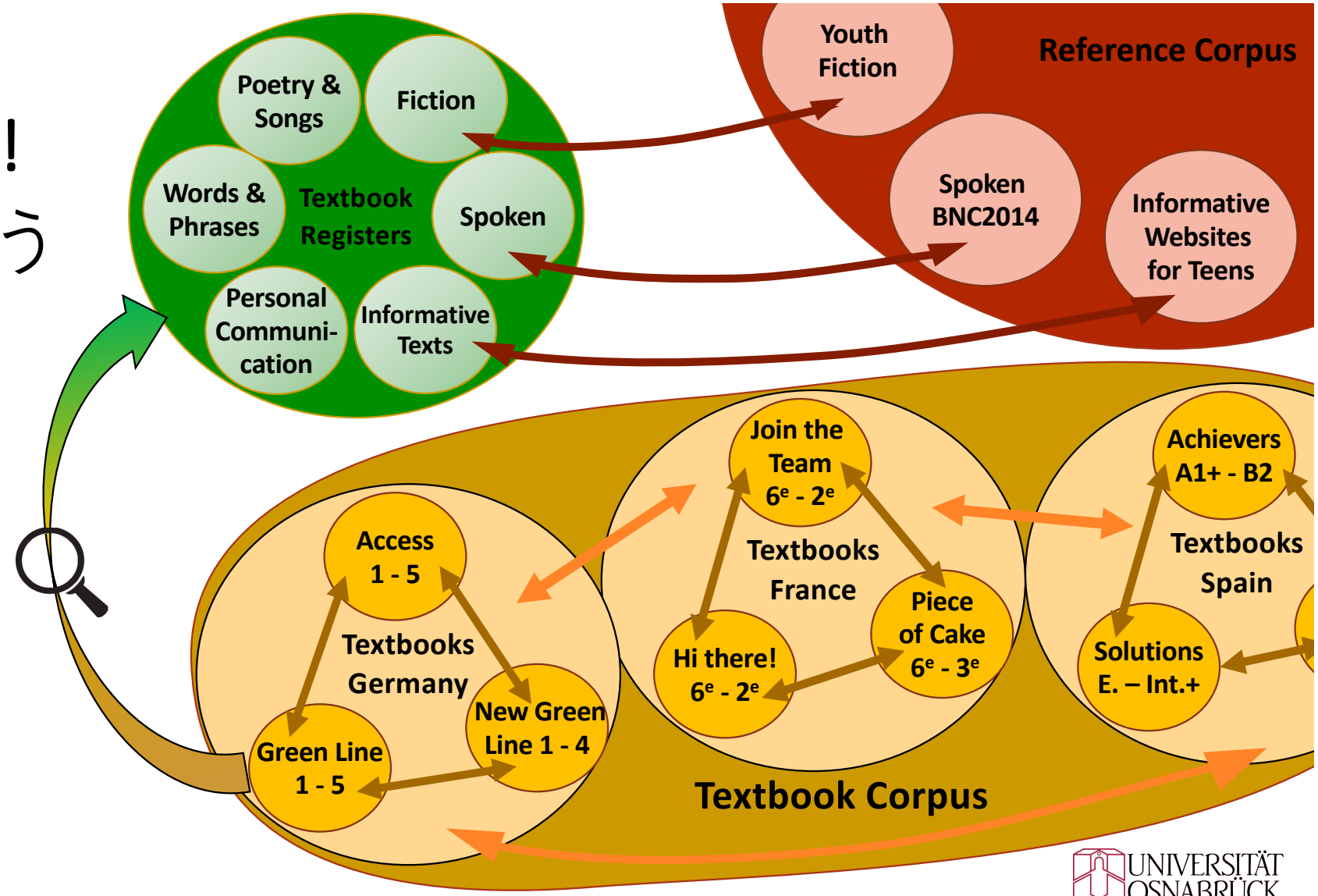
- Textbook Corpora → representative of language input in EFL classroom
- Corpus size, representativeness and balance
- Mark-up & annotation
- Reference corpora for comparison

Thank you!
ありがとう

Elen Le Foll

elefoll@uos.de

@ElenLeFoll



UNIVERSITÄT OSNABRÜCK

References I

BARBIERI, F., & ECKHARDT, S. E. (2007). Applying corpus-based findings to form-focused instruction: The case of reported speech. *Language Teaching Research*, *11*(3), 319–346.

BIBER, D., CONRAD, S., & CORTES, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

BIBER, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

BIBER, D., (2006). University language: a corpus-based study of spoken and written registers, Studies in corpus linguistics. Amsterdam: John Benjamins.

CULLEN, R., & KUO, I.-C. (2007). Spoken grammar and ELT course materials: a missing link? *Tesol Quarterly*, *4*(2), 361–386.

GABRIELATOS, C. (2006). Corpus-based evaluation of pedagogical materials: If-conditionals in ELT coursebooks and the BNC. In *Paper presented at the 7th Teaching and Language Corpora Conference.*

KOPROWSKI, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*, *59*(4), 322–332.

LE FOLL, E. (2019). *MAKING Explicit: The Use of MAKE in School English Textbooks*. Presented at the IVACS 2019, Dortmund.

LE FOLL, E. (2020). *Exploring the Registers of School EFL Textbooks Using Multi-Dimensional Analysis*. Presented at the ICAME41, Heidelberg.

LE FOLL, E. (forthcoming). *Textbook English: A Corpus-Based Analysis of the Language of EFL textbooks used in Secondary Schools in France, Germany and Spain.* Osnabrück University.

LJUNG, M. (1990). *A study of TEFL vocabulary*. Stockholm: Almqvist & Wiksell International.

LJUNG, M. (1991). Swedish TEFL meets reality. In S. Johansson & A.-B. Stenström (Eds.), *English Computer Corpora: Selected Papers and Research Guide* (pp. 245–256).

References II

MEUNIER, F., & GOUVERNEUR, C. (2007). The treatment of phraseology in ELT textbooks. In *Corpora in the foreign language classroom* (pp. 119–139).
MEUNIER, F., & Gouverneur, C. (2009). New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (Ed.), *Studies in Corpus Linguistics* (Vol. 33, pp. 179–201). Amsterdam: John Benjamins.
MINDT, D. (1987). *Sprache, Grammatik, Unterrichtsgrammatik: Futurischer Zeitbezug im Englischen I*. Frankfurt: Diesterweg.
RÖMER, U. (2004). A corpus-driven approach to modal auxiliaries and their didactics. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 185–199). Amsterdam: John Benjamins.
RÖMER, U. (2005). *Progressives, Patterns, Pedagogy.* Amsterdam: John Benjamins.
SINCLAIR, J., & RENOUF, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140–158). Harlow: Longman.
TESCH, F. (1990). *Die Indefinitpronomina* some *und* any *im authentischen englischen Sprachgebrauch und in Lehrwerken*. Tübingen: Narr.
TONO, Y. (2004). Multiple Comparisons of IL, L1 and TL Corpora: The Case of L2 Acquisition of Verb Subcategorization Patterns by Japanese Learners of English. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Studies in Corpus Linguistics* (Vol. 17, pp. 45–66). Amsterdam: John Benjamins.
Usó-Juan, E., & Martínez-Flor, A. (2010). The teaching of speech acts in second and foreign language instructional contexts. In *Pragmatics across Languages and Cultures* (pp. 423–442). Berlin: Walter de Gruyter.
VELLENGA, H. (2004). Learning Pragmatics from ESL & EFL Textbooks: How Likely? *TESL-EJ Teaching English as a Second or Foreign Language*, *8*(2), n. p.
VINE, E. W. (2013). Corpora and coursebooks compared: Category ambiguous words (Vol. 1, pp. 463–478). LCR 2011, Presses universitaires de Louvain.

Images

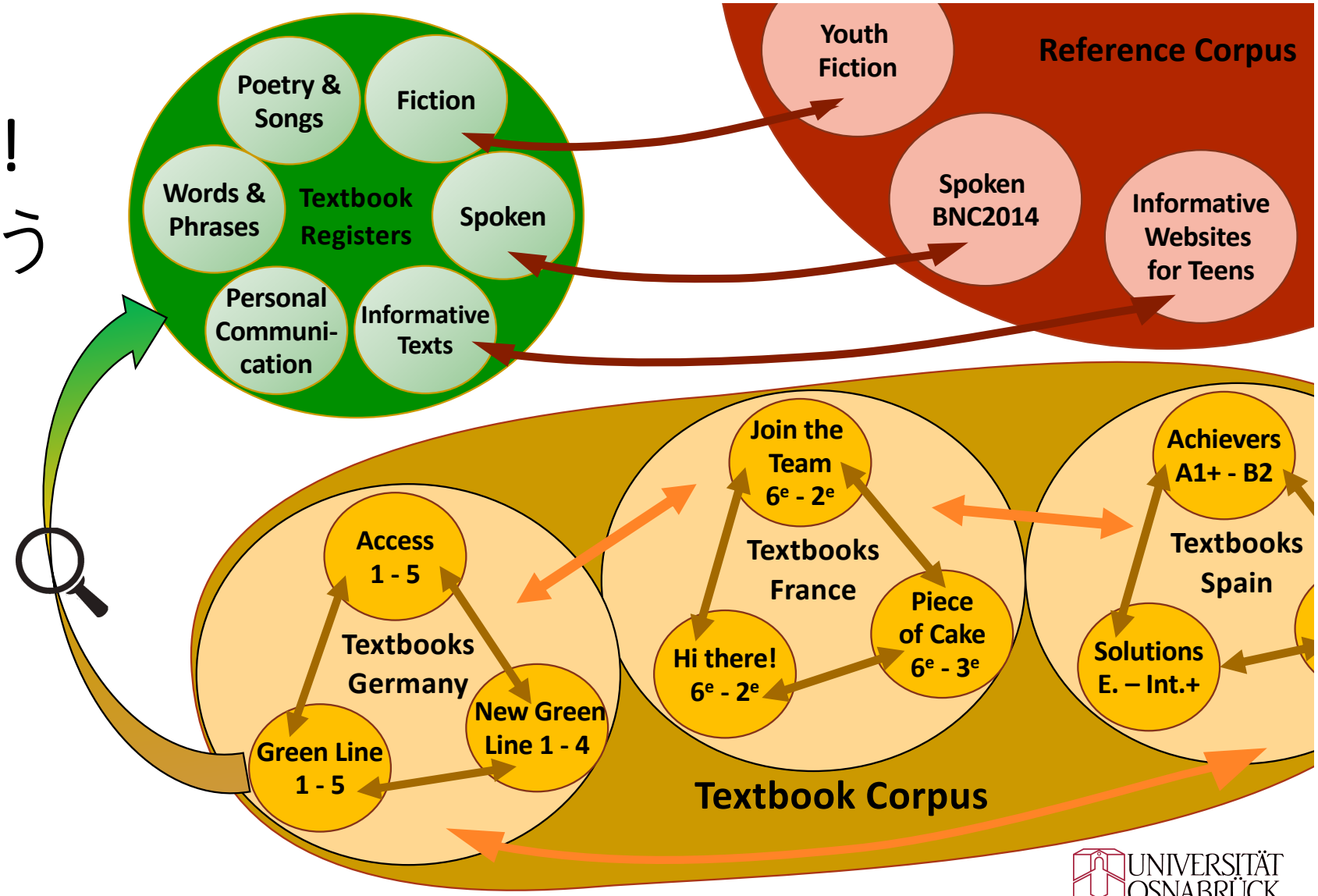
- Textbook examples from Green Line (Klett): https://klettbib.livebook.de/978-3-12-547140-5/
- Textbook examples from Piece of Cake (Le livre scolaire): https://www.lelivrescolaire.fr/page/14192789
- All cliparts from https://publicdomainvectors.org
- All plots by Elen Le Foll CC-BY 4.0 
- Palettes for R plots from {suffrager}

Thank you!
ありがとう

Elen Le Foll

elefoll@uos.de

@ElenLeFoll



UNIVERSITÄT OSNABRÜCK