Information Density and Fragments in German

SFB 1102/ Project B3

Eva Horch, Robin Lemke, Ingo Reich
i.reich@mx.uni-saarland.de

IDeaL
SFB 1102

UNIVERSITÄT
DES
SAARLANDES

1.
General Overview

Fragments in German

Fragments

Fragments are construed as antecedentless (overtly) non-sentential utterances that convey propositional content and illocutionary force.

Headlines:

8. März 2015, 17:29 "Tatort" aus Wien

Jeder Mensch ein Kauz

(SZON, 09.03.2015)

Fall Teresa Z.

Richterin degradiert Prügelpolizisten

(SZON, 09.03.2015)

Fragments in German

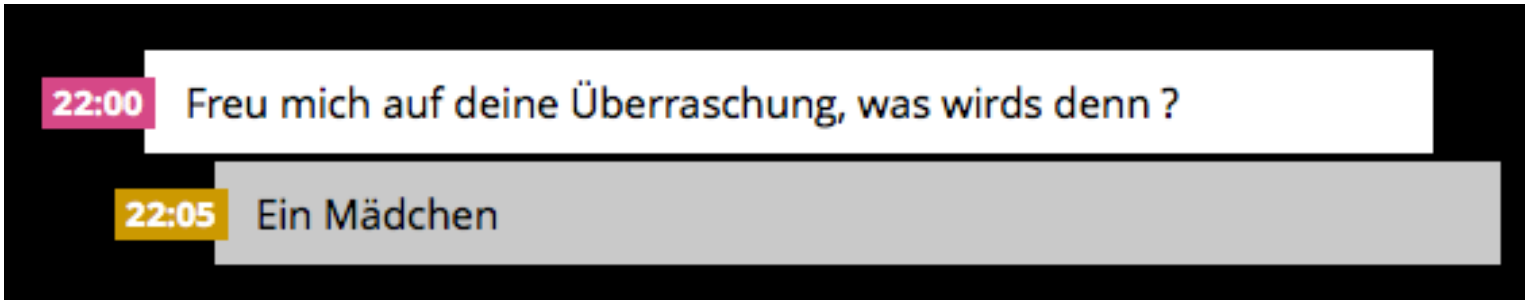
Advertisements:

Zeit für eine neue Form.
Der neue CLS AMG Shooting Brake.

Formsprache: Beeindruckend. Der CLS Shooting Brake fasziniert mit […]

Mercedes-Benz.
Das Beste oder nichts.

SMS:



Main Objectives

Theoretical foundations:

- exhaustive classification of fragments
- syntactic, semantic and pragmatic modelling

Corpus studies:

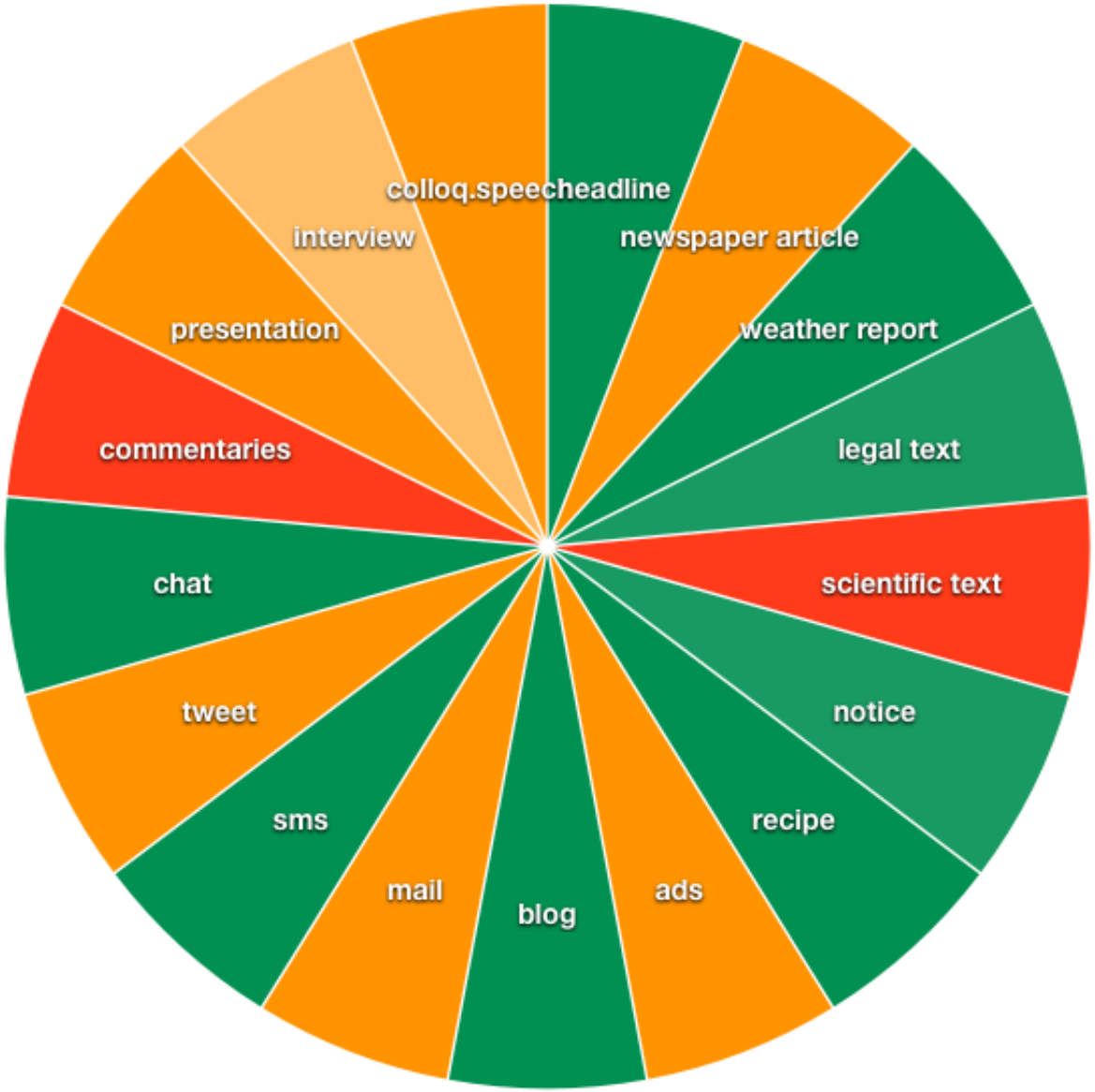
- build corpora of different text types (2.000 sentences each)
- quantitative and qualitative analysis of fragments in those corpora
- analysis of density profiles of those corpora using language models

Psycholinguistic experiments:

- conduct psycholinguistic experiments (based on corpus data), e.g. by
- measuring surprisal of a fragment relative to different text types

2.
Corpus Studies

Fragments & text type



Text types – debit and credit

- ✓ **headlines**: bild.de, spiegel-online.de, sueddeutsche.de, zeit.de (6 days, 2012)
- ✓ **recipes**: ZDF Topfgeldjäger & Küchenschlacht, daskochrezept.de, (kochrezepte.de, petitchef.de)
- ✓ **weather reports**: allewetter.de, dwd.de (Deutscher Wetterdienst), donnerwetter.de, wetter.de, wetter.info, wetternet.de, wetteronline.de (5 days in dec., 2012)
- ✓ **legal texts**: german basic law, Hochschulrahmengesetz, rental contract
- ✓ **sms**: SMS corpus (Werbesprache.net)
- ✓ **blogs**: blogger.de (4 different authors)
- ✓ **chat**: Dortmunder ChatKorpus
- ✓ **notice**: contact ad (Kontaktanzeigen.de, locanto.de), classifieds (rheinpfalz.de: car mart, job market, eviction market, personal announcements)

Text types – debit and credit

□/✓	**presentation**:	Baum corpus 2014 (SR1/2/3, Unser Ding)
□/✓	**tweets**:	Zenner corpus 2013
□/✓	**interview**:	SWR1 “Leute”
□/✓	**ads**:	canvassing of 2013, Scherer’s Corpus (cars, beauty, lifestyle)
□/✓	**mails**:	∅
□/✓	**colloquial speech**:	TueBa-DS
□/✓	**newspaper articles**:	TueBa-DZ
□	**commentaries:**	∅
□	**scientific texts:**	∅

Workflow

(Ideal) workflow:

1. collect raw data ($\geq$ 2.000 sentences per text type)
2. POS-tagging, tree-tagging to allow for search routines
3. annotation of different fragment types on an additional layer (*ideally* on a server with a multi-user interface)
4. make corpora available in well-established tools (e.g. ANNIS)

Open questions:

1. What kind of database is best used? (*apart from Excel* ...)
2. What kind of taggers seem reasonable *wrt* elliptical [!] corpus data?
3. What kind of annotation sofware provides a multi-user interface
4. *and* furthermore easily converts data into e.g. ANNIS-format?

Language Models: Core Questions

1. Do different text types differ *wrt* their *density profile*, i.e., is for example an SMS corpus denser than a newspaper corpus?
2. Do those differences covary with the use of (different) fragments, i.e., are fragments systematically used to modulate *density profiles*?
3. *Probably related to this point:* Does the predictability of a fragment depend on the text type it occurs in or just on the local context?
4. Are there text type-characteristic fragments / density profiles?
5. What else could LMs tell us about fragments & text type?

Example: Object Drop

Recipes (Topfgeldjäger, 03.03.15):

In jede Muschel einen großen Esslöffel Nussbutter, ein wenig Chili, einige Scheiben Zitronengras, zwei Scheiben Knoblauch, einige Thymianblättchen und eine kleine Prise Salz geben. Anschließend die Deckel auf die Muscheln setzen, den Rand mit Blätterteig verschließen und mit etwas Eigelb einpinseln. Die Muscheln möglichst gerade auf das vorgeheizte Salz im Backofen setzen und bei 210 °C Umluft acht Minuten garen.

Anschließend herausnehmen, auf Teller setzen und in der Schale servieren.

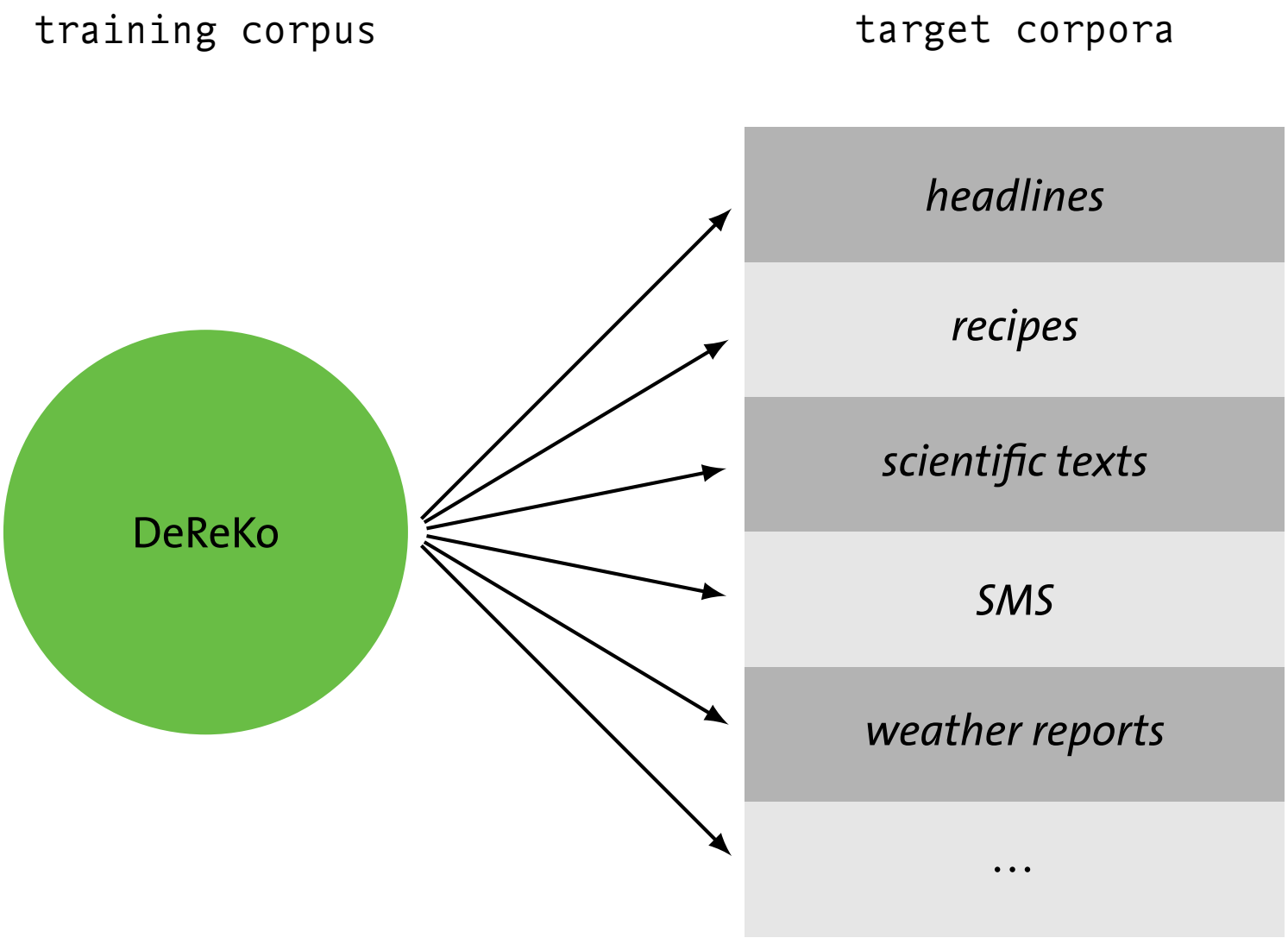
Headlines (SZON, 09.03.15):

IS im Irak

Erst zerstören, dann plündern

- *Is object drop characteristic for recipes?*
- *Does object drop modulate the density profile of recipes?*
- *Is object drop triggered by troughs in the density profile?* …

Language Models



3. Experiments

Follow ups (Selection)

- Does the none-use of text type characteristic fragments (e.g. object drop in recipes) affect in any systematic way reading times?
- Does the use of not text type characteristic fragments (e.g. copula ellipsis in spoken register) affect in any systematic way reading times?
- Does the acceptability of fragments covary with the text types?
- Does ungrammaticality play any systematic role (e.g. article drop)?

...

We are grateful for any suggestions!