Compiling a diverse web corpus for South Tyrolean German - STirWaC

Sarah Schulz, Verena Lyding, Lionel Nicolas

sarah.schulz@ugent.be
LT^3, Language and Translation Technology Team
Ghent University

{verena.lyding;lionel.nicolas}@eurac.edu
Institute for Specialised Communication and Multilingualism
European Academy of Bolzano

July 22, 2013

Outline

State of the art

Web-based corpora

- large web-based corpora for national varieties of several languages available (cp. eg. Roth (2012), Baroni et al. (2009), Cook and Hirst (2012))
- BooTCaT by Baroni and Bernardini (2004) tool which facilitates the compilation of web-based corpora
- corpus building for minority languages - web crawling software by Scannell (2007)

Problems of state-of-the art approaches

Quantity, quality and restriction

State-of-the-art approaches assume 2 main criteria...

- ... a certain variety has its own top-level domain
- ... a domain contains enough content to build a large corpus

But a lot of small varieties do not meet these criteria.

Our main contributions

Compiling web-based corpora for smaller varieties

In the following we ...

- ... explain a procedure for web-based corpora of language varieties that are not restricted to one single-top level domain and face data sparsity (example: STirWaC: corpus of South Tyrolean German).
- .... introduce a procedure for improving the balance of the corpus in terms of the diversity of texts
- ... describe and evaluate the resulting STirWaC, the largest ever-built web-corpus for South Tyrolean German

Overview of the method

restriction

harvest a base corpus

quantity

crawling a larger corpus

quality

expanding the coverage over less represented text types

Overview of the method

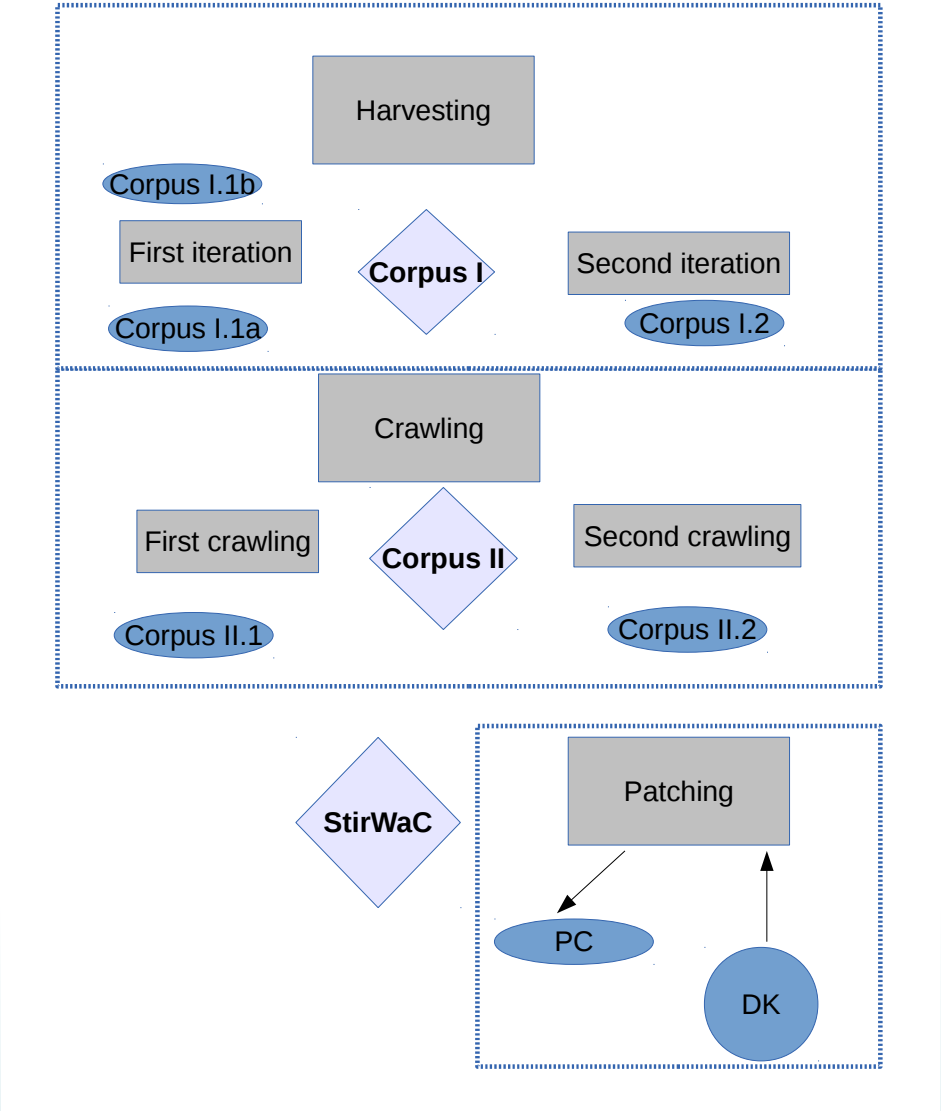


Figure: Work flow

Overview of the method

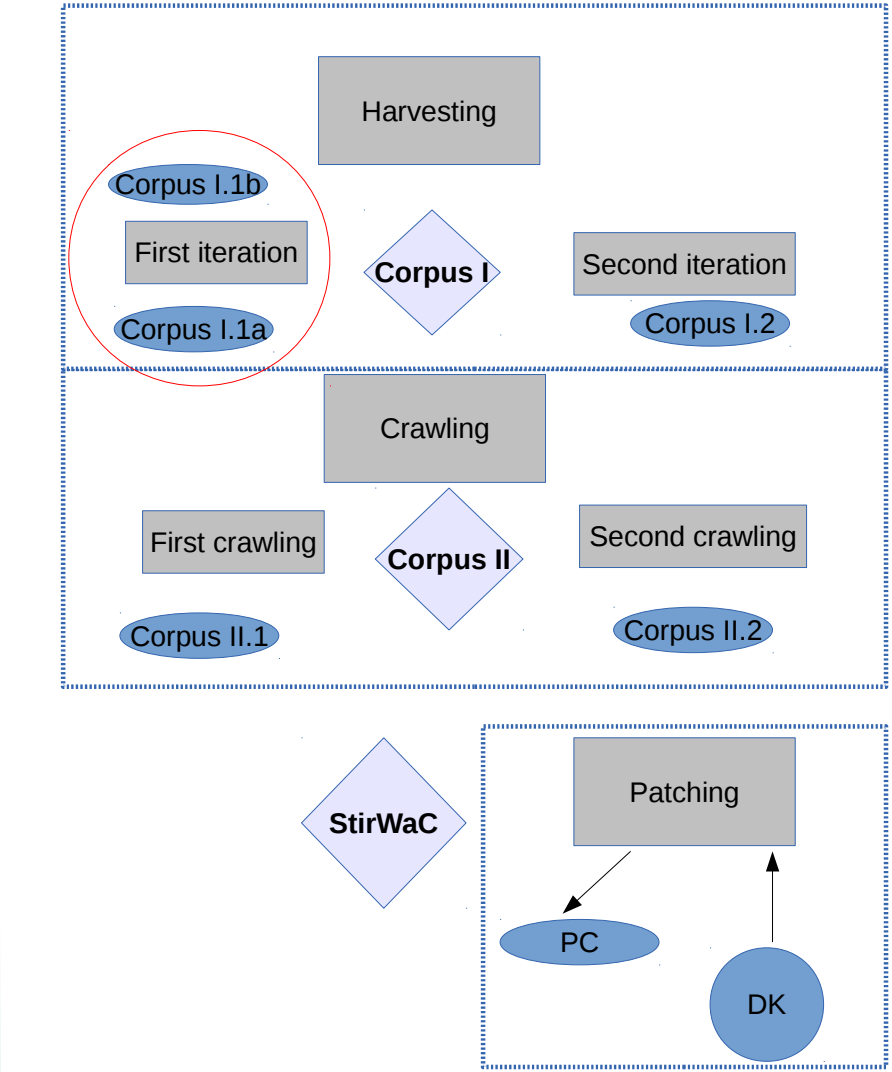


Figure: Work flow: Harvesting

	Corpus	**I.1a**	**I.1b**	**I.1**	**I.2**	**I**
	Method	*Harvesting*	*Harvesting*	I.1a ∩ I.1b	*Harvesting*	I.1 ∩ I.2
Setup	Domains	.it	¬{.de}	-	all	-
	Seeds	100 terms	42 terms	-	1,000 terms	-
	Search Tuples	500 of length 3	500 of length 2	-	5,000 of length 2	-
	Max Results/Query	50	50	-	30	-
	Upper Limit	25,000	25,000	15,060	150,000	40,588
Results	Unique URLs	15,572	10,420	14,930	103,896	39,813
	DeDuper-ed Docs	11,070	3,990	14,869	25,719	39,502
	Tokens	9,658,731	4,108,360	13,442,536	39,405,480	50,734,333
	Lemmas	109,200	70,255	123,255	196,479	210,657

Table: Summary of corpus I.

Overview of the method

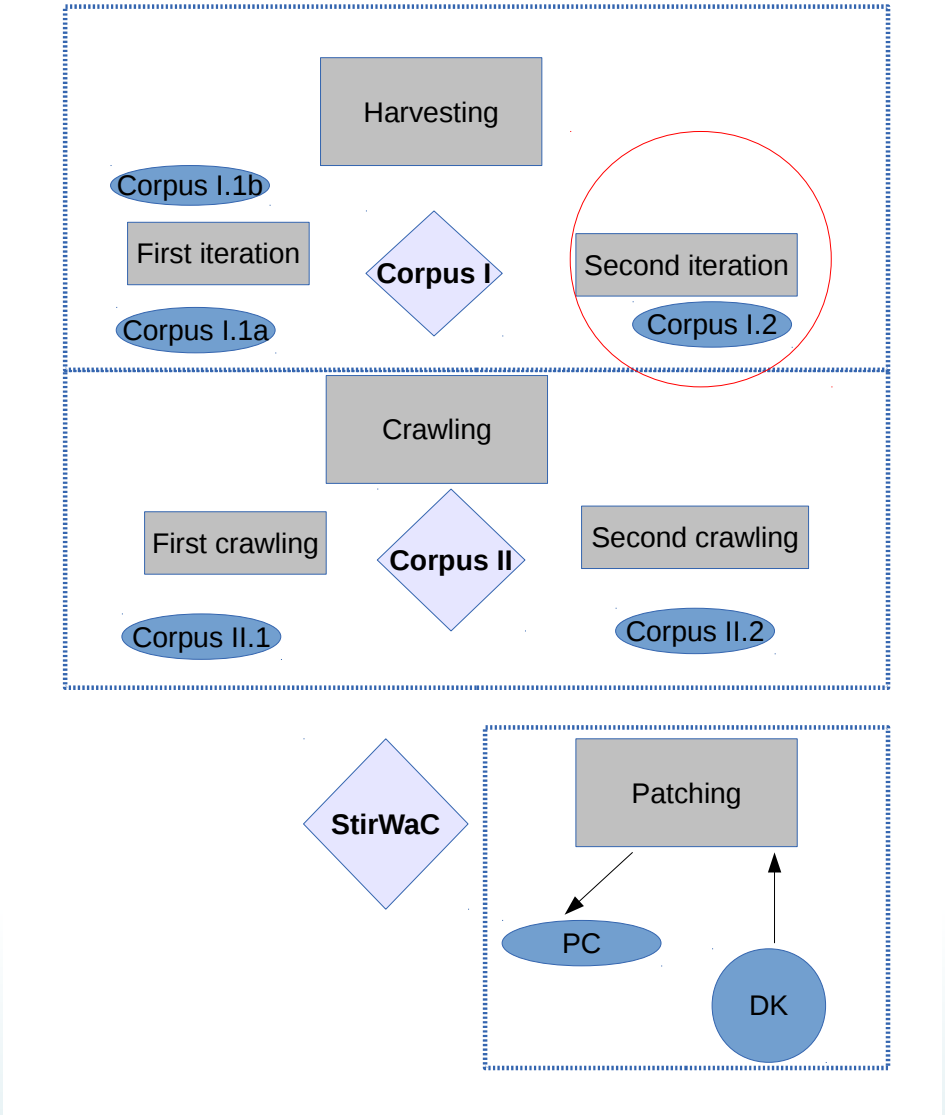


Figure: Work flow: Harvesting

	Corpus	**I.1a**	**I.1b**	**I.1**	**I.2**	**I**
	Method	*Harvesting*	*Harvesting*	I.1a ∩ I.1b	*Harvesting*	I.1 ∩ I.2
Setup	Domains	.it	¬{.de}	-	all	-
	Seeds	100 terms	42 terms	-	1,000 terms	-
	Search Tuples	500 of length 3	500 of length 2	-	5,000 of length 2	-
	Max Results/Query	50	50	-	30	-
	Upper Limit	25,000	25,000	15,060	150,000	40,588
Results	Unique URLs	15,572	10,420	14,930	103,896	39,813
	DeDuper-ed Docs	11,070	3,990	14,869	25,719	39,502
	Tokens	9,658,731	4,108,360	13,442,536	39,405,480	50,734,333
	Lemmas	109,200	70,255	123,255	196,479	210,657

Table: Summary of corpus I.

Distribution of top-level domains

Domain \ Corpus	*I*.1*a*	*I*.1*b*	*I*.1	*I*.2	*I*
.it	11,070 (100.0%)	1,256 (31.48%)	12,149 (81.71%)	3,551 (13.81%)	15,099 (38.22%)
.de	-	-	-	10,544 (41.00%)	10,544 (26.70%)
.at	-	373 (9.35%)	373 (2.51%)	2,779 (10.81%)	3,090 (7.82%)
.ch	-	126 (3.16%)	125 (0.84%)	989 (3.85%)	1,102 (2.79%)
other	-	2,235 (56.02%)	2,222 (14.94%)	7,856 (30.55%)	9,667 (24.47%)
total	11,070	3,990	14,869	25,719	39,502

Table: Distribution of top-level domains of harvested corpora

Overview of the method

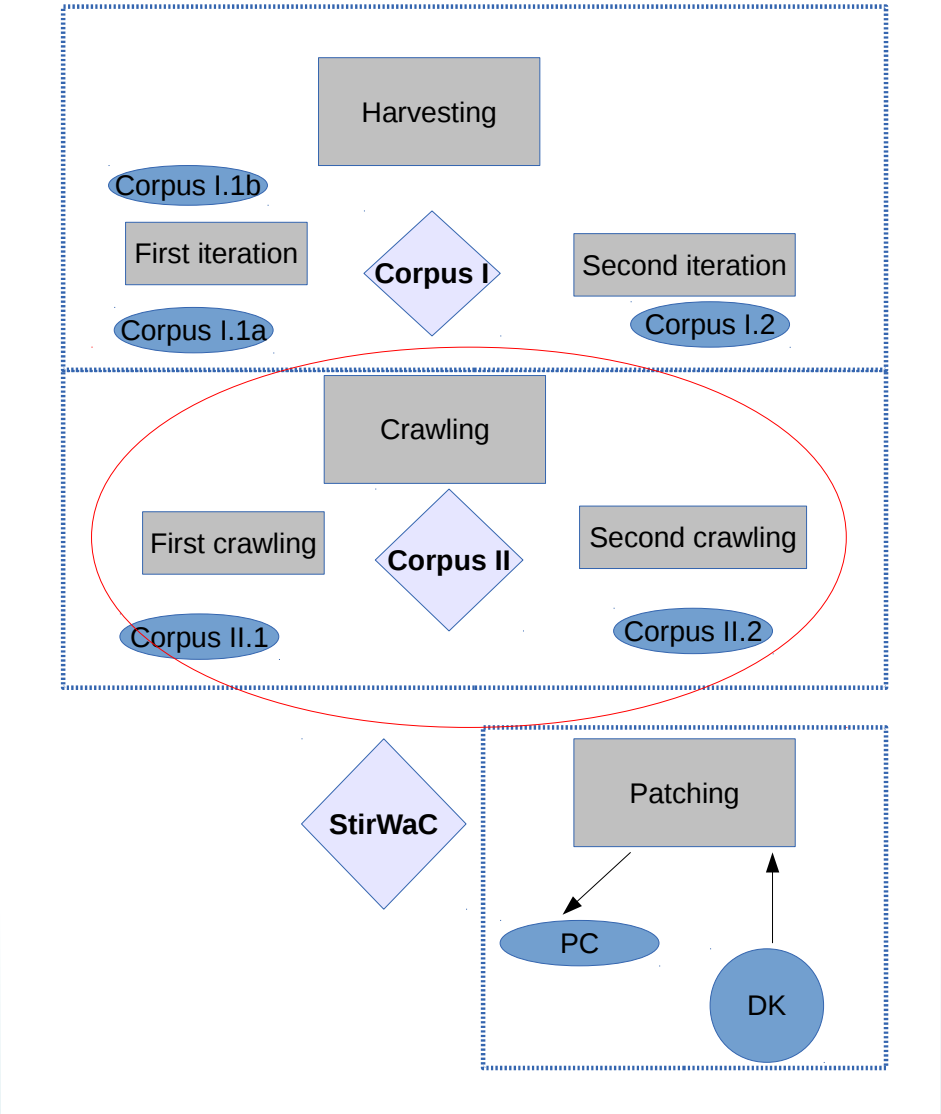


Figure: Work flow: Crawling

Distribution of top-level domains

Domain \ Corpus	*I*	*II*.1	*II*.2	*II*	**STirWaC**
.it	15,099 (38.22%)	30,573 (66.63%)	4,027 (17.26%)	32,759 (51.25%)	36,561 (42.15%)
.de	10,544 (26.70%)	723 (1.58%)	537 (2.30%)	1,171 (1.83%)	11,668 (13.45%)
.at	3,090 (7.82%)	116 (0.25%)	145 (0.62%)	215 (0.34%)	3,283 (3.78%)
.ch	1,102 (2.79%)	75 (0.16%)	30 (0.13%)	104 (0.16%)	1,204 (1.39%)
other	9,667 (24.47%)	14,401 (31.38%)	18,597 (79.69%)	29,674 (46.42%)	34,033 (39.23%)
total	39,502	45,888	23,336	63,923	86,749

Table: Distribution of top-level domains.

Summery of all corpora

	Corpus	I	II.1	II.2	II	STirWaC
	Method	*Harvesting*	*Crawling*	*Crawling*	II.1 ∩ II.2	I ∩ II
Setup	Domains	-	I.1 \{.at, .ch}	I.2[1]\{.de, .at, .ch}	-	-
	Seeds	-	14,245[2] URLs	4,625 URLs	-	-
	Search Tuples	-	-	-	-	-
	Max Results/Query	-	-	-	-	-
	Upper Limit	40,588	-	-	69,224	103,425
Results	Unique URLs	39,813	135,285	65,554	64,892	88,651
	DeDuper-ed Docs	39,502	45,888	23,336	63,923	**86,749**
	Tokens	50,734,333	29,777,384	22,170,902	47,869,771	**82,262,840**
	Lemmas	210,657	160,035	157,264	195,981	237,623

Table: Summary of the corpus.

[1] From these URLs only the single shortest URL per site was kept.
[2] This should be 14,371 but our exclusion pattern was a tad too generous.

Overview of the method

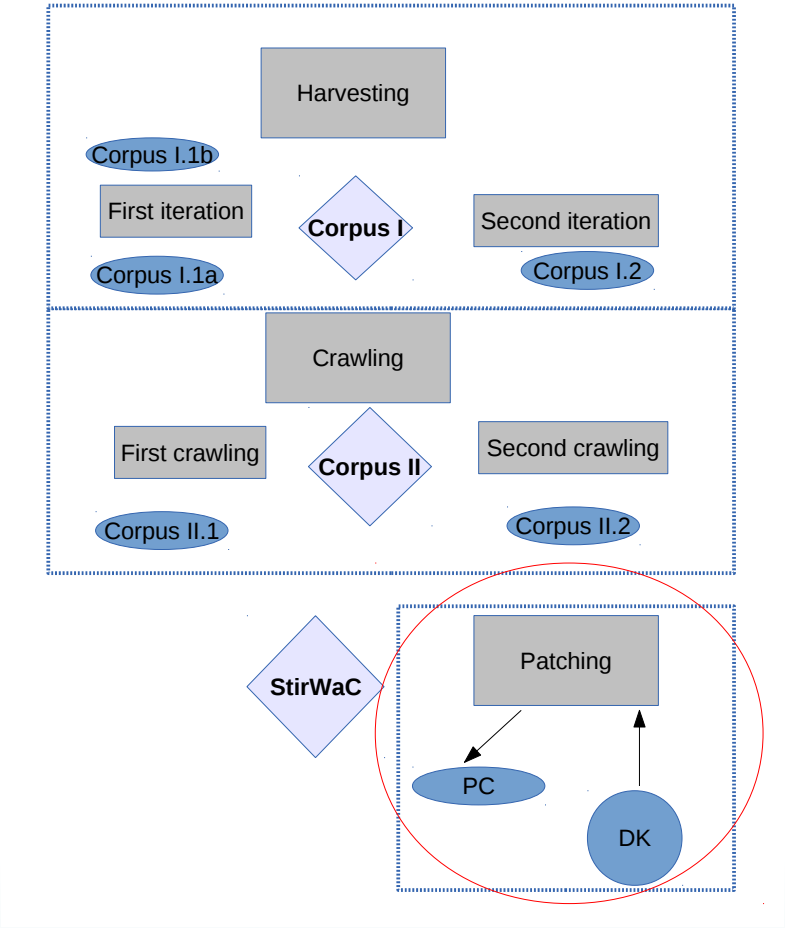


Figure: Work flow: Patching

Patching to increase diversity

Assessing corpus diversity and text types

- patching the STirWaC corpus with documents not reached by standard BootCaT harvest and crawling.
- reach a better balancedness in terms of text type
- text type: texts that have a high similarity to each other with respect to a bunch of features

Patching to increase diversity

Basic idea

a specialized seed term list, specific to subcorpora of certain text types, can be used to detect and exploit previously missed parts of the Internet.

Tasks to tackle

- group the text into subcorpora as basis for seed term extraction → left to future work
- classify our documents according to text features
- verify that seed term list compiled from grouped subcorpora enables us to retrieve documents from the same text type

Patching to increase diversity

Underlying approach

- method developed by Forsyth and Sharoff (2013)
- manually evaluated text set on several linguistic aspects
- attributes of texts used as coordinates of a vector
- attribute vectors are reduced to two and mapped on a 2D map
- plot STirWaC with the help of trained tool for standard German

Plotting texts on a 2D space with regard to their text features

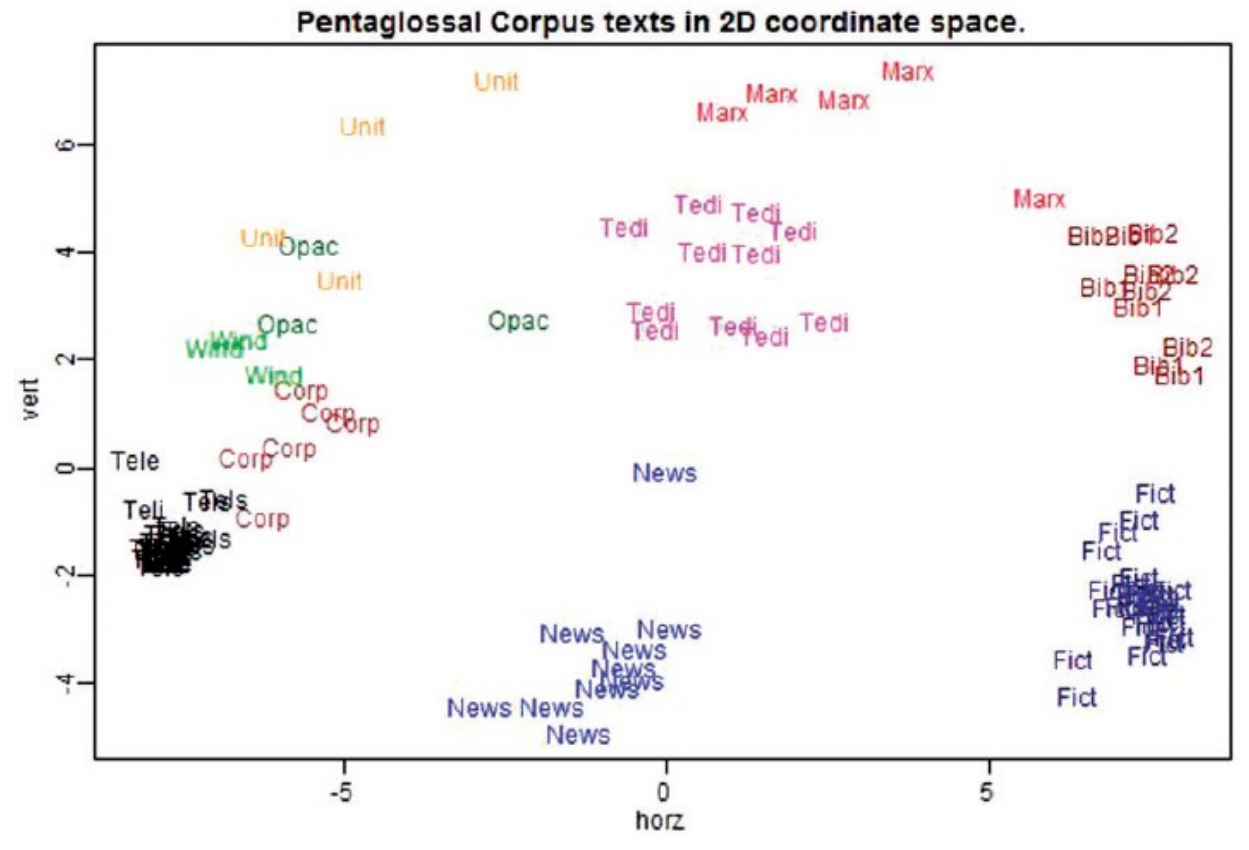
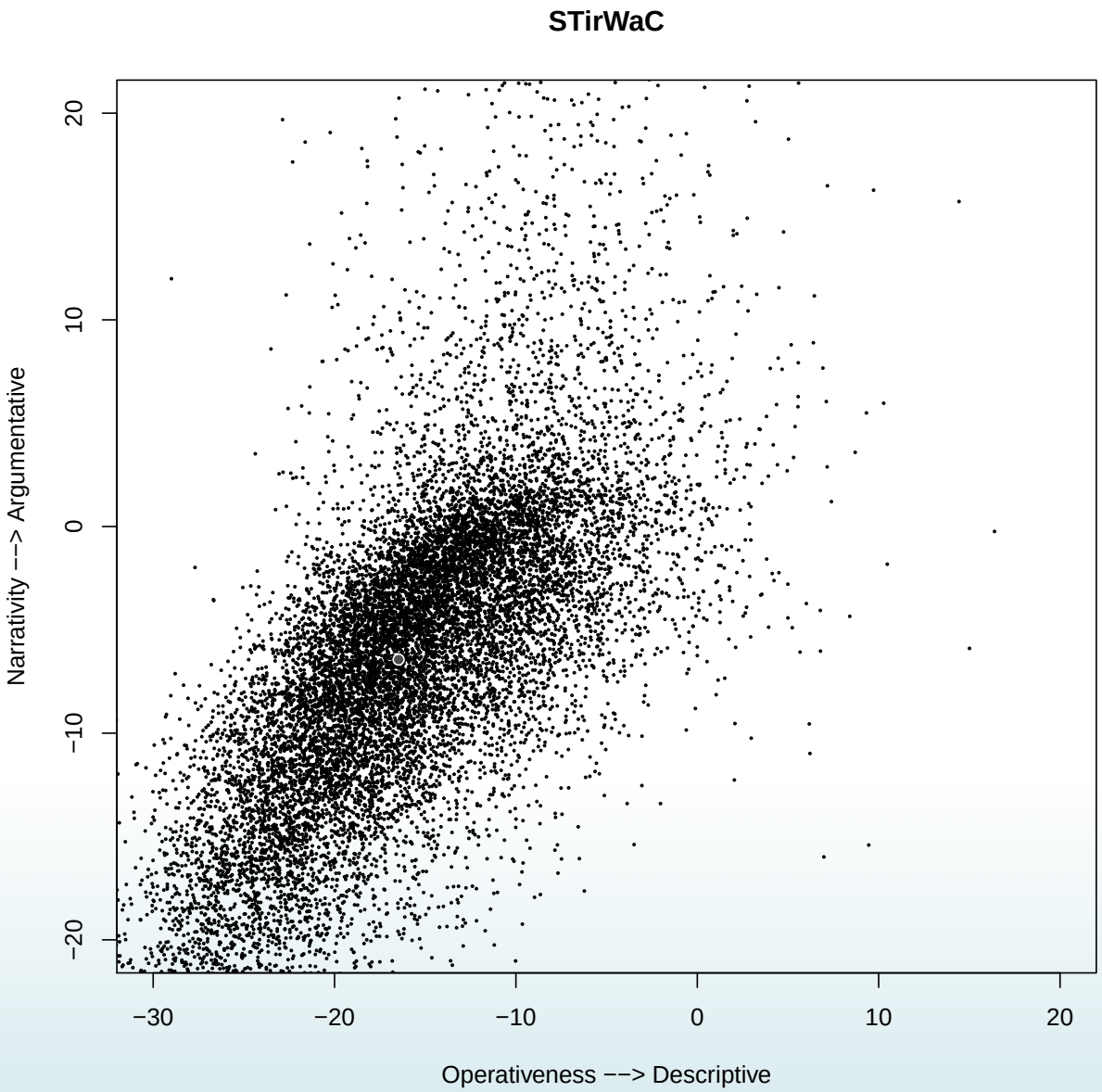
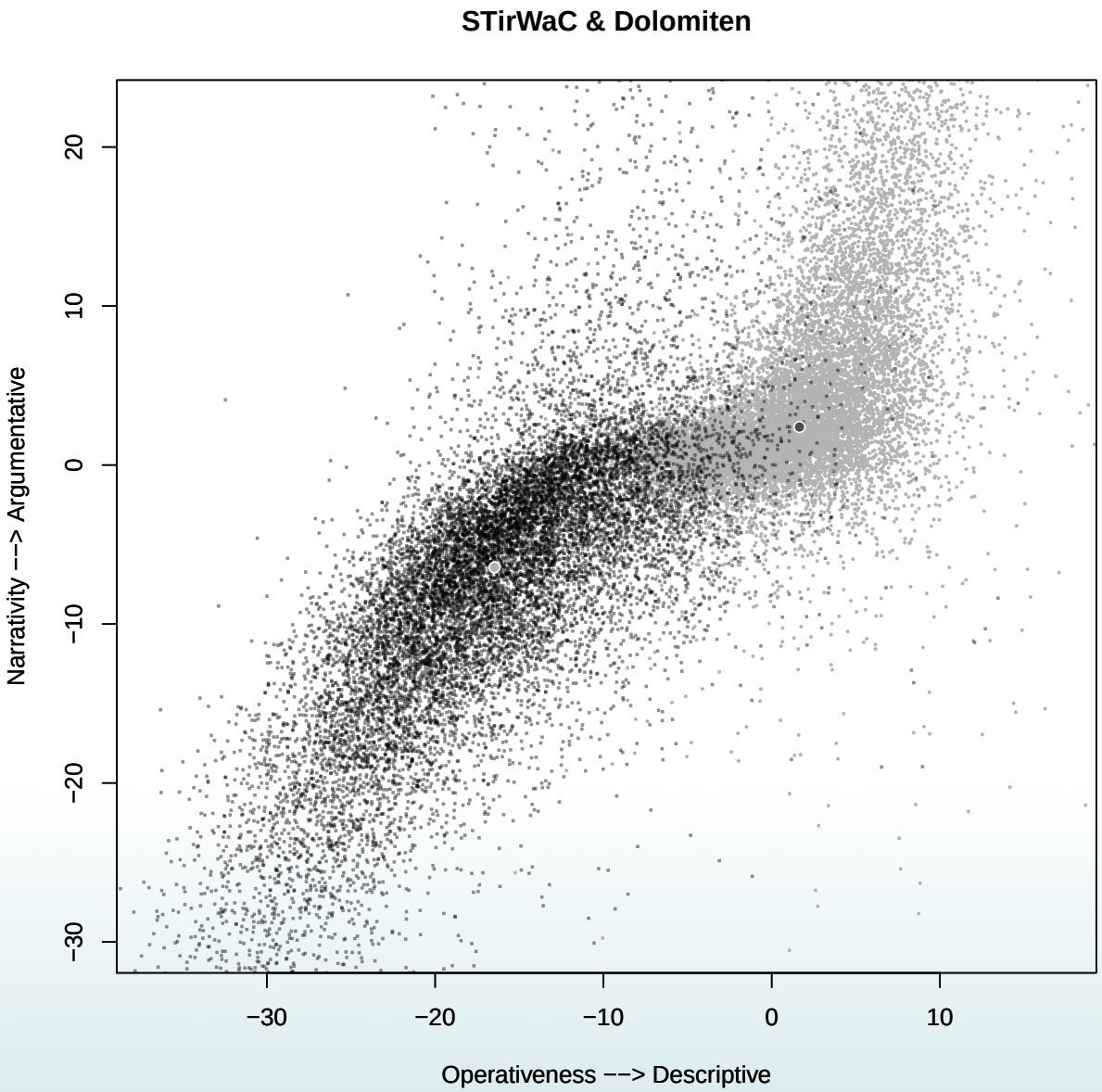


Figure: The pentaglossal corpus collected by Forsyth and Sharoff (2013) plotted on a 2D similarity space.

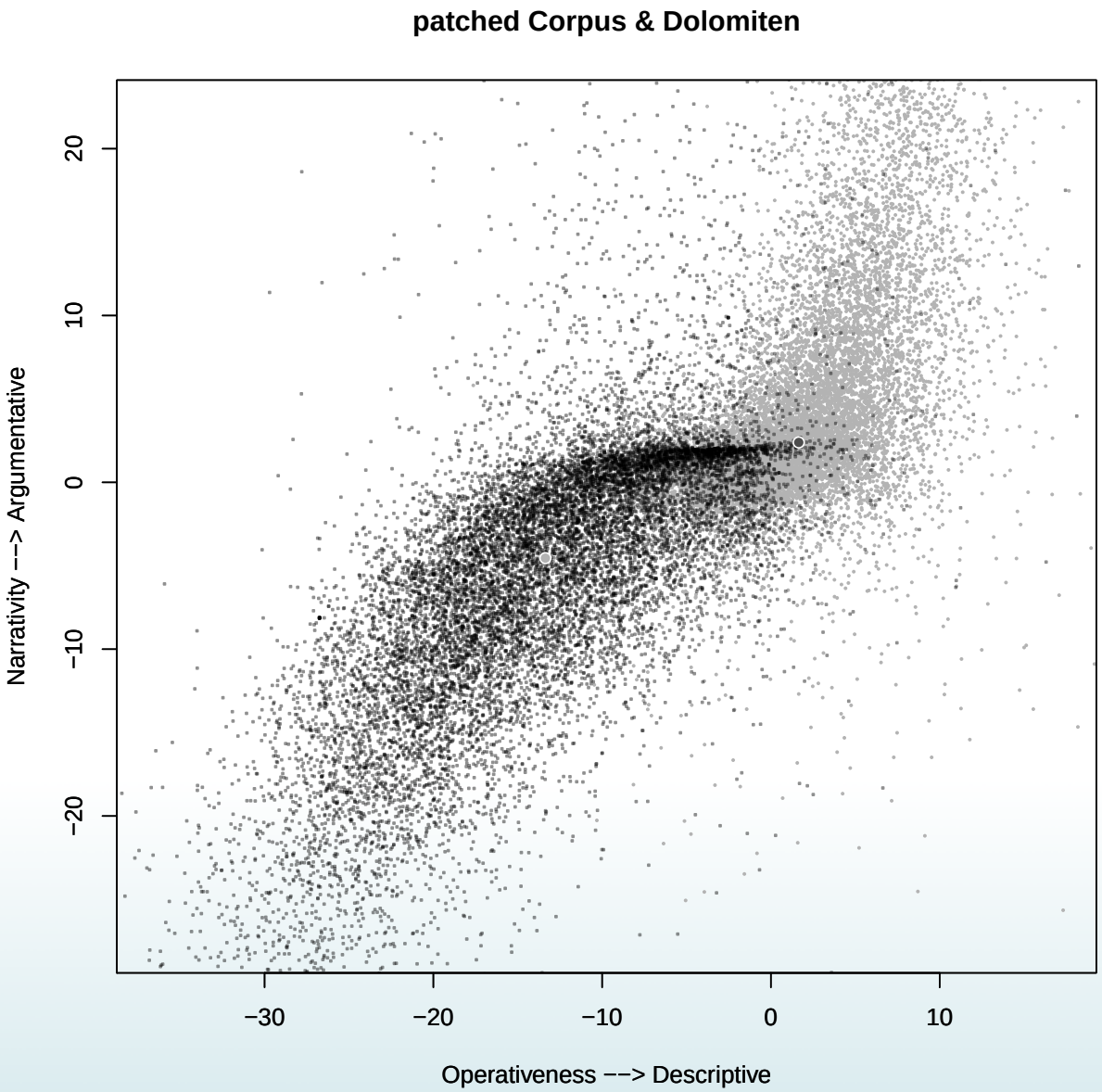
Plotting STirWaC



Filling the gap



Patching to increase diversity



Evaluation

collocation/term	Typical of	**rf**$_{at}$	**rf**$_{ch}$	**rf**$_{de}$	**rf**$_{st}$
wilder Knoblauch	*AT DE*	1.8	1.0	1.3	**4.9**
Blaulicht und Sirene	*CH DE*	2.2	**5.9**	3.7	2.4
Blaulicht und Folgetonhorn	*AT*	**4.0**	0	0	0
Blaulicht und Martinshorn	*DE*	1.8	1.5	**8.7**	0
in angetrunkenem Zustand	*CH DE*	0.7	**55.4**	2.0	37.7
Einspruch einlegen	*DE*	23.0	34.8	**90.8**	35.3
große Töne spucken	*DE*	**12.1**	9.8	11.8	0
Baukonzession	*STIR*	1.5	1.5	4.0	**305.1**
Handelsoberschule	*STIR*	0.4	0	0	**181.1**
Regionalrat	*STIR*	7.3	11.8	8.7	**494.8**
innerhalb <date>	*STIR*	0	0	0.3	**175.0**
halbmittag	*STIR*	0.4	0	0	**25.5**
weißer Stimmzettel	*STIR*	0	0	0	**6.1**

Table: Relative frequencies of characteristic n-grams over *STirWaC* (**rf**$_{st}$) and three other corpora covering documents in Austrian German (**rf**$_{at}$), Swiss German (**rf**$_{ch}$) and the standard German (**rf**$_{de}$) Roth (2012)

Conclusion

Conclusion

- we have built the largest South Tyrolean web corpus currently available
- corpus highly relevant for South Tyrolean German
- presented a blueprint approach for the compilation of specialized corpora of other language varieties
- introduced a new approach towards the extension of web corpora considering text type

Future work

- improve size and representativeness of STirWaC
- fully implement the grouping approach of subcorpora with respect to text type

Literature I

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *LREC*. European Language Resources Association.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Literature II

Cook, P. and Hirst, G. (2012). Do web corpora from top-level domains represent national varieties of english? In *Proceedings, 11th International Conference on Statistical Analysis of Textual Data / 11es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2012)*, pages 281–291, Liège.

Forsyth, R. S. and Sharoff, S. (2013). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing*.

Literature III

Roth, T. (2012). Using web corpora for the recognition of regional variation in standard german collocations. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*. Adam Kilgarriff and Serge Sharoff.

Scannell, K. P. (2007). The crbadn project: Corpus building for under-resourced languages.