The KFOR Text Corpus

Dr. Matthias Hecking
Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN)
Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE)
Abtl. Informationstechnik und Führungssysteme (ITF)
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
hecking@fgan.de

Content

- 1. Introduction
- 2. Corpora for Computational Linguistics
- 3. The KFOR Text Corpus
  - Token, Gazetteer
  - Named Entities
  - Verbal Group
- 4. Evaluation of the ZENON System
- 5. Conclusion

1 | Introduction - I

- Processing of human language as a critical capability in many future military applications (cf. [Steeneken, 1996]).
- Content analysis/extraction of free-form texts is important for any information operation of the Network Centric Warfare (NCW) concept (s. [NCW, 2001], p. 5-15).
- This can be realized through Information Extraction (IE) which is a natural language processing technique (cf. [Appelt, 1999], [Hecking, 2004a]).

1 | Introduction - II

- specific problem: content extraction of HUMINT reports
- ZENON project: The overall objective is to realize experimental systems for (partial) *content extraction of HUMINT reports* from the KFOR deployment of the Bundeswehr and to realize a possibility to evaluate the formal representation of the content.
- cf. [Hecking, 2004b], [Hecking, 2005a], [Hecking, 2006a], [Hecking, 2006b]
- For the realization of the IE module 4,498 English HUMINT reports are available.

1 | Introduction - III

- The efficiency of natural language processing systems must be evaluated.
- state of the art: comparison of the produced annotations (the extracted content) with the expected annotations
- The expected annotations are given by a corpus.
- (Text) Corpus = set of texts and associated annotations
- The text sort and the analysis objectives determine which syntactic and/or semantic annotations are needed.
- syntactic annotations = part-of-speech, conjugation information (e.g. 3rd pers sing), structure of nominal phrases (e.g. ART ADJ NOUN)...
- semantic annotations = name of cities, rivers, countries...

1 | Introduction - IV

- For the evaluation and improvement of the information extraction of the ZENON prototype the *KFOR Text Corpus* was realized.
- 4,498 HUMINT reports (mostly in English) from the KFOR deployment of the German Federal Armed Forces
- 800 of them manually annotated (= KFOR Corpus)
- The performance of the ZENON information extraction is quantitatively evaluated relative to the KFOR corpus.
- Since the KFOR corpus is classified, it is not freely available.
- The report (cf. [Hecking, 2006c]) is *not* classified.

2 | Corpora for Computational Linguistics - I

- Because we are not able to list all texts of a language variety (e.g. all HUMINT reports in English from 1980 to 2000) we have to build a sample of it.
- corpus for empirical research on written or spoken texts with annotations
- E.g.,
  - “The bomb did not ignite in the station of Koblenz.“
  - semantic annotation for the string “Koblenz”:

    city[40, 47, {name= Koblenz}]
  - i.e. this string is the name of a city and the name starts in position 40 and ends in 47.

2 | Corpora for Computational Linguistics - II

- corpora are
  - of finite size
  - very huge (but: micro-text corpus)
  - machine-readable
  - used as a standard reference
  - representative of the language variety
- examples
  - American National Corpus (ANC), over 20 million words (15.12.2005)
  - British National Corpus (BNC), 100 million words (2007)

2 | Corpora for Computational Linguistics - III

- different classes of annotation
  - textual/extra-textual: basic information, e.g., author name, date the text was written, the variety of the language, broad subject domain, ...
  - part-of-speech (POS): for each token; e.g. past participle, noun, adjective, ...
  - parsing: higher-level syntactic relationships, ‘treebanks’, e.g. ART ADJ NOUN
  - semantics: semantic relationship between entities, e.g. the AGENT of an action; semantic features of words
  - phonetic transcription: spoken language, phonemes
  - prosody: suprasegmental features of spoken language; e.g., stress, intonation, pauses, ...

3 | The KFOR Text Corpus - I

- 4,498 HUMINT reports (mostly in English) from the KFOR deployment of the Bundeswehr
- KFOR Corpus = 800 of them manually annotated
- 886,000 tokens; different annotation layers
- specialized micro-text corpus (cf. [McEnery, 2001])
- syntactic and semantic annotations
- first version produced automatically; corrected manually
- used tool: GATE (www.gate.ac.uk)
- formats: GATE-specific, GATE-specific in XML, ANC (American National Corpus) stand-off annotation, TIGER-XML

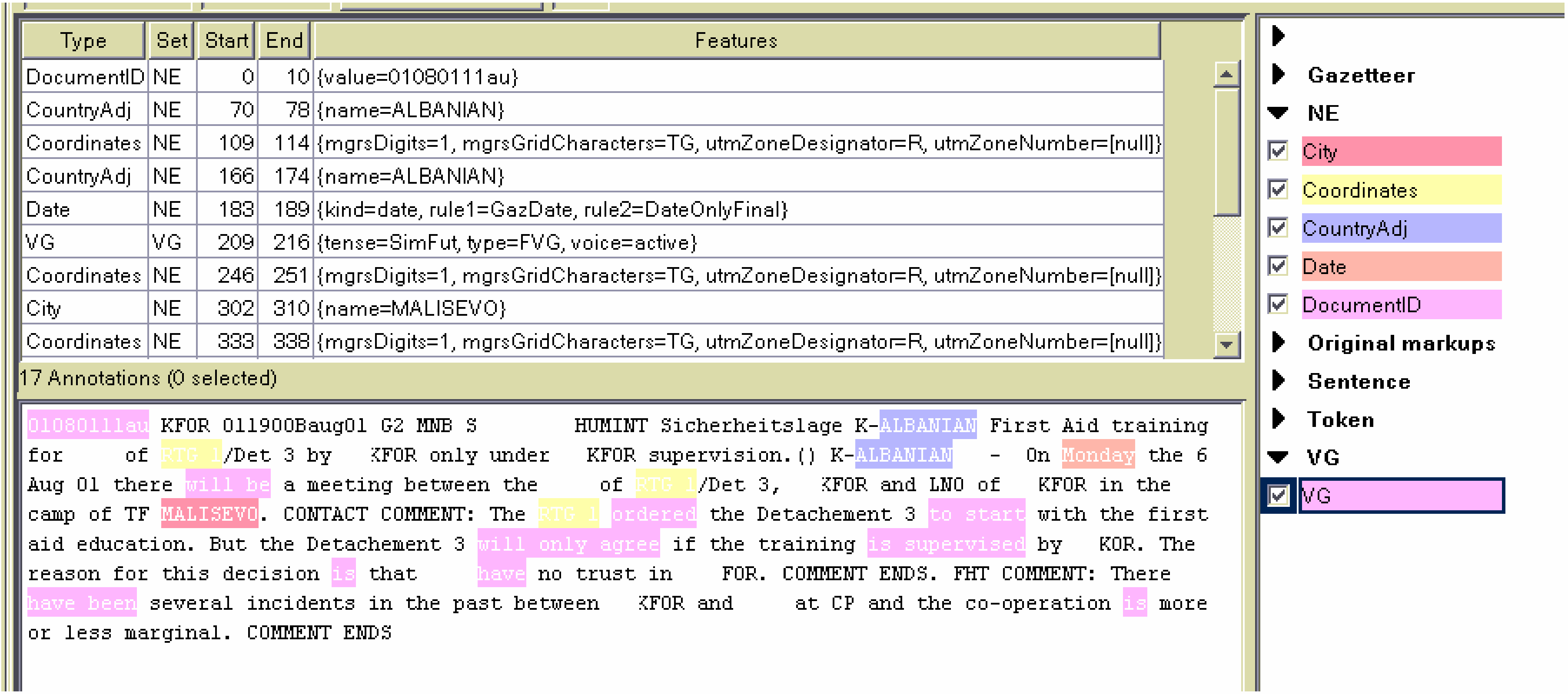
3 | The KFOR Text Corpus - II

- annotation layers:
  - Original markups: pre-formatted parts (e.g. addressee, topic, source)
  - Token: words, numbers, part-of-speech, lemma
  - Gazetteer: lists of names (e.g., first names, city names)
  - Sentence: sentences, begin and end of comments
  - Named entities (NE): names
  - Verb group (VG): verbal phrases
  - Thematic roles (ThRo): syntactic and semantic function of expressions in sentences (e.g., AGENT, TIME)

3 | The KFOR Text Corpus - III

Syntactical / semantical	Annotation layer	Annotation type	Checked manually
syntactical	Original markup	DocID, DTGMeldung, Einsatz, Empfaenger, Hauptthema, Koordinate, Meldung, Meldungstyp, Ort, Quelle, Sachverhalt, Schlagworte, Titel, Unterthema	no
syntactical	Token	Token, SpaceToken	no
semantical	Gazetteer	Lookup	no
syntactical	Sentence	Sentence Comment Split	yes yes no
semantical	NE	City, Company, Coordinates, Colour, CountryAdj, Currency, Date, DocumentID, GeneralOrg, MilDateTime, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time, Title	yes
syntactical	VG	VG	yes
semantical	ThematicRole	ThRo	yes

3 | The KFOR Text Corpus - IV



Type	Set	Start	End	Features
DocumentID	NE	0	10	{value=01080111au}
CountryAdj	NE	70	78	{name=ALBANIAN}
Coordinates	NE	109	114	{mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}
CountryAdj	NE	166	174	{name=ALBANIAN}
Date	NE	183	189	{kind=date, rule1=GazDate, rule2=DateOnlyFinal}
VG	VG	209	216	{tense=SimFut, type=FVG, voice=active}
Coordinates	NE	246	251	{mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}
City	NE	302	310	{name=MALISEVO}
Coordinates	NE	333	338	{mgrsDigits=1, mgrsGridCharacters=TG, utmZoneDesignator=R, utmZoneNumber=[null]}

17 Annotations (0 selected)

```
01080111au KFOR 011900Baug01 G2 MNB S        HUMINT Sicherheitslage K-ALBANIAN First Aid training
for      of RTG 1/Det 3 by   KFOR only under   KFOR supervision.() K-ALBANIAN    -  On Monday the 6
Aug 01 there will be a meeting between the      of RTG 1/Det 3,   KFOR and LNO of   KFOR in the
camp of TF MALISEVO. CONTACT COMMENT: The RTG 1 ordered the Detachement 3 to start with the first
aid education. But the Detachement 3 will only agree if the training is supervised by   KOR. The
reason for this decision is that      have no trust in    FOR. COMMENT ENDS. FHT COMMENT: There
have been several incidents in the past between   KFOR and     at CP and the co-operation is more
or less marginal. COMMENT ENDS
```

3 | The KFOR Text Corpus: Token

- basic building blocks
- words, numbers, etc.
- types: SpaceToken, Token
- e.g.: “KFOR”

Token Token xxx yyy {category= NNP, kind=word, length=4, orth=allCaps, root=kfor, string= KFOR}

Type	Feature name	Feature value
Token	affix	String
	category	CC, CD, DT, EX, FW, IN, JJ, JJR, JJS, JJSS, -LRB-, LS, MD, NN, NNP, NNPS, NNS, NP, NPS, PDT, POS, PP, PRPR$, PRP, PRP$, RB, RBR, RBS, RP, STAART, SYM, TO, UH, VBD, VBG, VBN, VBP, VB, VBZ, WDT, $’, ’-’, ’(’, ’.’,
	kind	word, number, symbol, punctuation
	length	Number
	orth	allCaps, lowercase, mixedCaps, upperInitial
	position	startpunct, endpunct
	root	String
	string	String

3 | The KFOR Text Corpus: Gazetteer

- expressions identified through lists of names (so-called gazetteers)
- used for the production of other annotations
- features: majorType, minorType
- e.g.: “BERLIN”

Type	majorType	minorType
Lookup	colour	<no>
	country_adj	<no>
	date	day, month
	location	city, country, province, river, region
	number	<no>
	organization	general, military, political, company
	person_first	female, male
	time	ampm, hour, zone
	title	civilian, police, military, male, female

Lookup Gazetteer xxx yyy {majorType=location, minorType=city}

3 | The KFOR Text Corpus: Named Entities - I

Type	Feature name	Feature value
MilitaryOrg	name	String

- most extensive: 20 types
- e.g.: national, supra-national and non-governmental military entities are treated as military organizations
- e.g.: “091100Bjul01”

Type	Feature name	Feature value
MilDateTime	year month day hour minute timeZone	String 1, ..., 12 1, ..., 31 1, ..., 24 1, ..., 60 UTC, ...

MilDateTime NE xxx yyy {year=01, month=7, day=9, hour=11, minute=0, timeZone=B}

3 | The KFOR Text Corpus: Named Entities - II

- problems with the semantic annotation
- words can to be polysem (more than one meaning)
- language users connote (pos. or neg.) words differently
- E.g., "KPC" (Kosovo Protection Corps)
  - official view of the political institutions (cf. [UNMIK, 2006]): a kind of THW (German Federal Agency for Technical Relief) ➡ PoliticalOrg
  - another opinion: the KPC as a successor of the Kosovo Liberation Army (KLA) is a terrorist organization ➡ MilitaryOrg
- rules: official view of the political institutions, view used by most language users, annotator decides (set of defined rules)

3 | The KFOR Text Corpus: Verbal Group - I

- verbal expressions
- e.g.: “CPC *can no more tolerate* this ladys behavior.”

Type	Feature name	Feature value
VG	adverb adverbPost infinitive negation	String String String yes
	special	HadBetter, SupposedTo, BeTo, HaveTo, GotTo, GoingTo, AbleTo, UnableTo, UsedTo
	tense	BeVBG, BeVBN, FutCon, FutPer, FutPerCon, HaveVBG, HaveVBN, HaveBeenVBG, Inf, Pas, PasCon, PasPer, [illegible] PerCon, Pre, [illegible] PrePerCon, SimFut, SimPas, SimPre
	type	FVG, MODAL, NFVG, PART, SPECIAL
	voice	active, passive

VG xxx yyy {adverb=more, infinitive=tolerate, modal=can,
neg=yes, type=MODAL, voice=active}

3 | The KFOR Text Corpus: Verbal Group - II

- problems
  - VGs can also be part of a nominal phrase (NP); e.g., "check-up" in "a hardware check-up of the planned test“
  - non-native English speakers use intelligible words which are not in the dictionary; e.g., the verb "to unclarify“
  - a verb complex can be divided into parts; one complex or two? e.g., in "Should they have a coalition?"

VG xxx yyy {modal=should, type=MODAL}

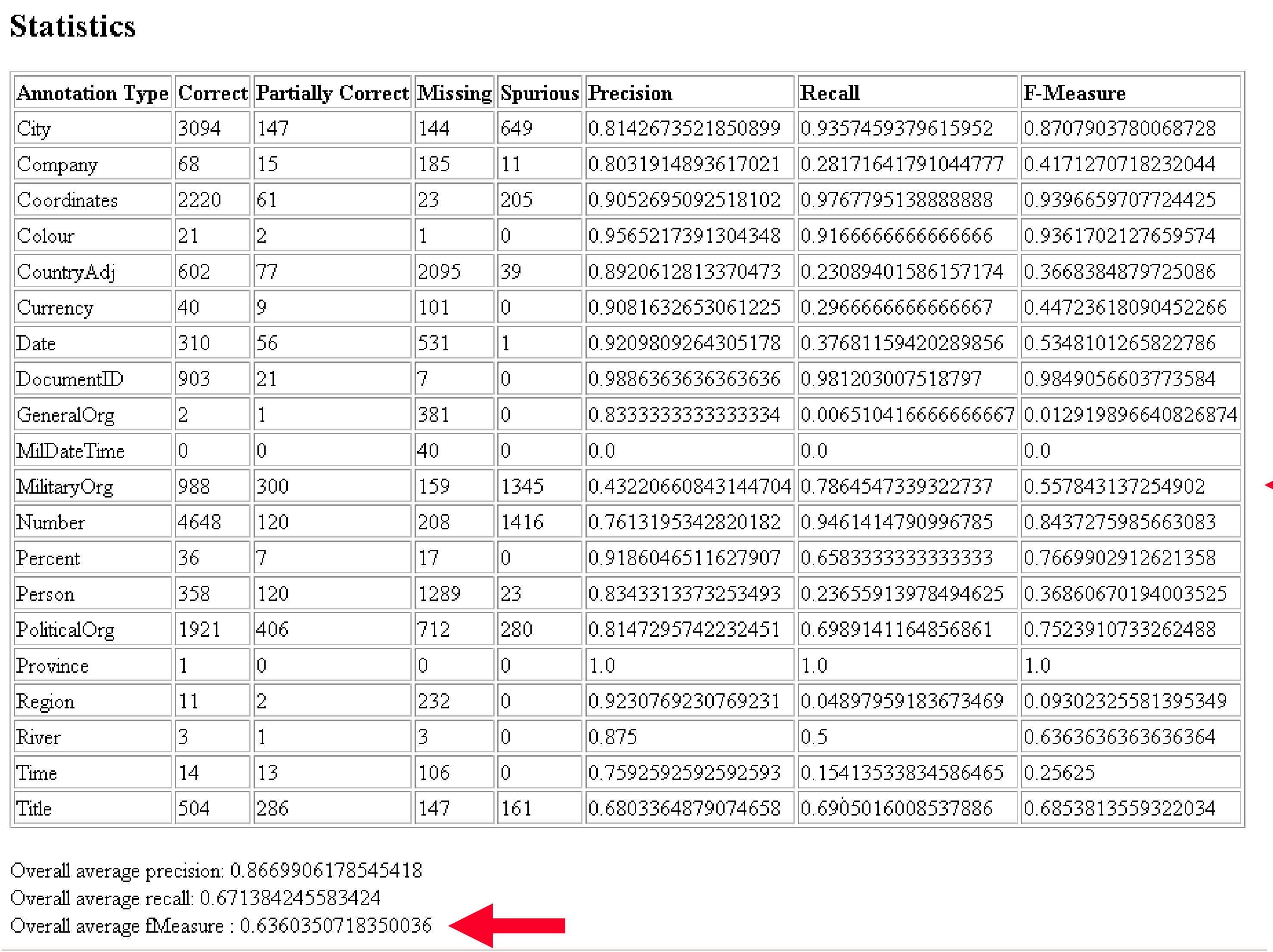
VG xxx yyy {infinitive=have, tense=Inf, type=NFVG, voice= active}

4 | Evaluation of the ZENON System - I

- The KFOR corpus was used to evaluate the information extraction component of the ZENON system.
- metrics:
  - Precision P: the number of correctly identified items as a percentage of the number of *all* items identified
  - Recall R: the number of correctly identified items as a percentage of the total number of *correct* items
  - F-measure: weighted average of the Precision and Recall
- "Corpus Benchmark Evaluation Tool” (GATE): to compare two different sets of annotations on the same documents
- next slide: all NE annotations produced by ZENON compared with the those of the KFOR corpus; all 800 documents; 12/2006 vs. 5/2007

4 | Evaluation of the ZENON System - II

(all) NE annotations, all docs, 12/2006



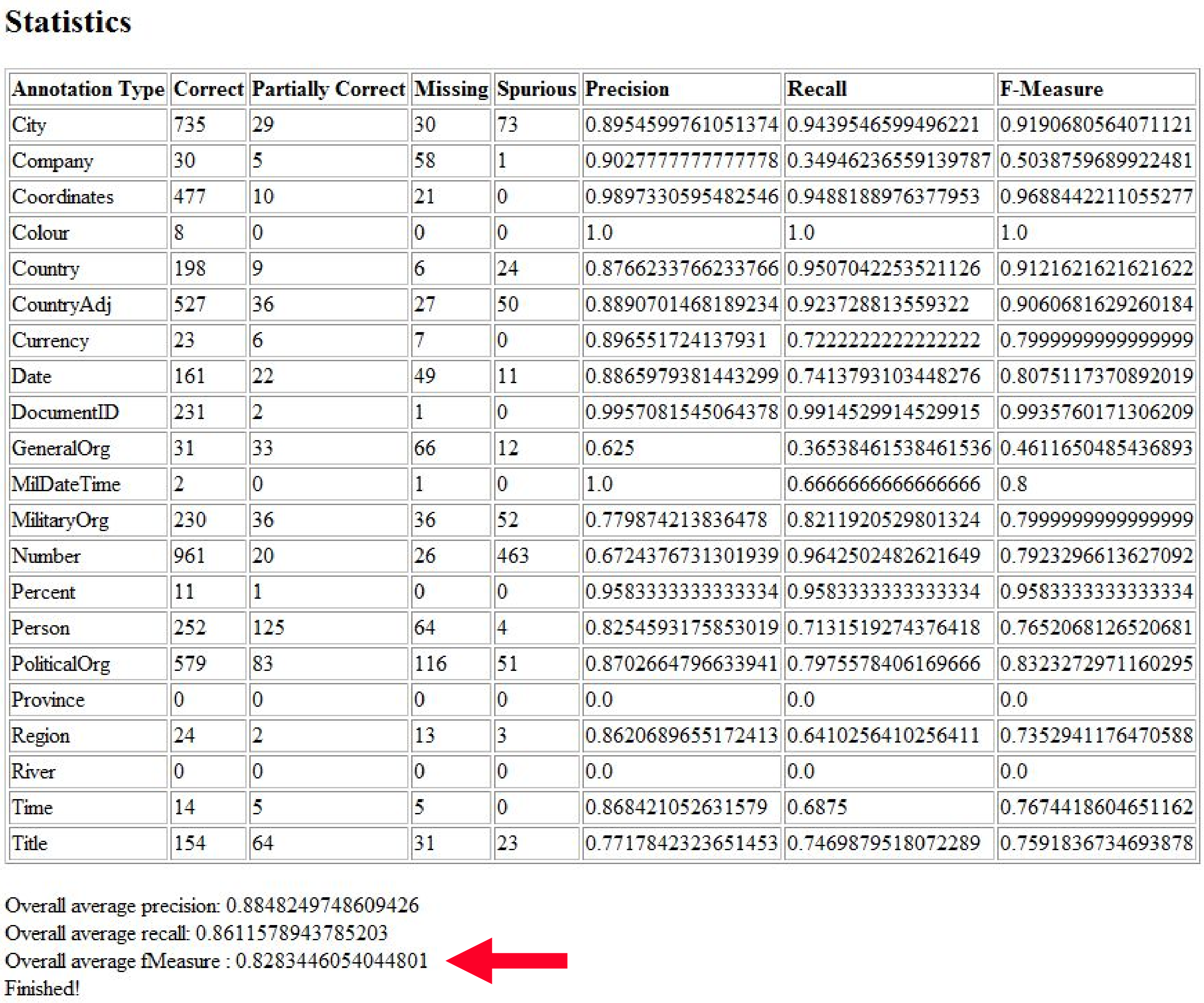
Statistics

Annotation Type	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
City	3094	147	144	649	0.8142673521850899	0.9357459379615952	0.8707903780068728
Company	68	15	185	11	0.8031914893617021	0.2817164179104477	0.4171270718232044
Coordinates	2220	61	23	205	0.9052695092518102	0.9767795138888888	0.9396659707724425
Colour	21	2	1	0	0.9565217391304348	0.9166666666666666	0.9361702127659574
CountryAdj	602	77	2095	39	0.8920612813370473	0.23089401586157174	0.3668384879725086
Currency	40	9	101	0	0.9081632653061225	0.2966666666666667	0.44723618090452266
Date	310	56	531	1	0.9209809264305178	0.37681159420289856	0.5348101265822786
DocumentID	903	21	7	0	0.9886363636363636	0.981203007518797	0.9849056603773584
GeneralOrg	2	1	381	0	0.8333333333333334	0.006510416666666667	0.012919896640826874
MilDateTime	0	0	40	0	0.0	0.0	0.0
MilitaryOrg	988	300	159	1345	0.43220660843144704	0.7864547339322737	0.557843137254902
Number	4648	120	208	1416	0.7613195342820182	0.9461414790996785	0.8437275985663083
Percent	36	7	17	0	0.9186046511627907	0.6583333333333333	0.7669902912621358
Person	358	120	1289	23	0.8343313373253493	0.23655913978494625	0.36860670194003525
PoliticalOrg	1921	406	712	280	0.8147295742232451	0.6989141164856861	0.7523910733262488
Province	1	0	0	0	1.0	1.0	1.0
Region	11	2	232	0	0.9230769230769231	0.04897959183673469	0.09302325581395349
River	3	1	3	0	0.875	0.5	0.6363636363636364
Time	14	13	106	0	0.7592592592592593	0.15413533834586465	0.25625
Title	504	286	147	161	0.6803364879074658	0.6905016008537886	0.6853813559322034

Overall average precision: 0.8669906178545418
Overall average recall: 0.671384245583424
Overall average fMeasure : 0.6360350718350036

4 | Evaluation of the ZENON System - III

all NE annotations, all docs, 5/2007



Statistics

Annotation Type	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
City	735	29	30	73	0.8954599761051374	0.9439546599496221	0.9190680564071121
Company	30	5	58	1	0.9027777777777778	0.34946236559139787	0.503875968922481
Coordinates	477	10	21	0	0.9897330595482546	0.9488188976377953	0.9688442211055277
Colour	8	0	0	0	1.0	1.0	1.0
Country	198	9	6	24	0.8766233766233766	0.9507042253521126	0.9121621621621622
CountryAdj	527	36	27	50	0.8890701468189234	0.923728813559322	0.9060681629260184
Currency	23	6	7	0	0.896551724137931	0.7222222222222222	0.7999999999999999
Date	161	22	49	11	0.8865979381443299	0.7413793103448276	0.8075117370892019
DocumentID	231	2	1	0	0.9957081545064378	0.9914529914529915	0.9935760171306209
GeneralOrg	31	33	66	12	0.625	0.36538461538461536	0.461165048543689З
MilDateTime	2	0	1	0	1.0	0.6666666666666666	0.8
MilitaryOrg	230	36	36	52	0.779874213836478	0.8211920529801324	0.7999999999999999
Number	961	20	26	463	0.6724376731301939	0.9642502482621649	0.7923296613627092
Percent	11	1	0	0	0.9583333333333334	0.9583333333333334	0.9583333333333334
Person	252	125	64	4	0.8254593175853019	0.7131519274376418	0.7652068126520681
PoliticalOrg	579	83	116	51	0.8702664796633941	0.7975578406169666	0.8323272971160295
Province	0	0	0	0	0.0	0.0	0.0
Region	24	2	13	3	0.8620689655172413	0.6410256410256411	0.7352941176470588
River	0	0	0	0	0.0	0.0	0.0
Time	14	5	5	0	0.868421052631579	0.6875	0.7674418604651162
Title	154	64	31	23	0.7717842323651453	0.7469879518072289	0.7591836734693878

Overall average precision: 0.8848249748609426
Overall average recall: 0.8611578943785203
Overall average fMeasure : 0.8283446054044801
Finished!

4 | Evaluation of the ZENON System - IV

- The information extraction functionality of the ZENON system was improved.
- ca. 15 new transducer, updated lists, improved transducer
- overall improvements:
    - overall average P: 0.87 ➡ 0.88
    - overall average R: 0.67 ➡ 0.86
    - overall average F-measure: 0.64 ➡ 0.83
- specific improvements in F-measure:
    - CountryAdj: 0.37 ➡ 0.91
    - Person: 0.37 ➡ 0.77
- specific degradation in F-measure:
    - Number: 0.84 ➡ 0.79

5 | Conclusion

- introduction; why we need a corpus for the ZENON project
- corpora for empirical research in Computational Linguistics
- the KFOR text corpus
    - annotation layers
    - token, gazetteer
    - named entities
    - verbal group
- The KFOR corpus was used to evaluate the information extraction component of the ZENON system.