Collocation Extraction Based on Syntactic Criteria

Violeta Seretan

Department of Translation Technology
Faculty of Translation and Interpreting
University of Geneva

September 2013

Recent Advances in Natural Language Processing, 7-13 September 2013, Hissar, Bulgaria
Acknowledgment

Language Technology Laboratory, Department of Linguistics, University of Geneva
Acknowledgment

Language Technology Laboratory, Department of Linguistics, University of Geneva

Eric Wehrli  Luka Nerima  Paola Merlo
1. On collocations
2. Importance of collocations
3. Distiguishing features
4. The need for morphosyntactic analysis
5. Syntax-based extractors
6. Method, results, evaluation
7. Conclusion
I chose to **run for the presidency** at this moment in history because I believe deeply that we cannot solve the challenges of our time unless we solve them together.

Source: https://my.barackobama.com/page/content/hisownwords/
I chose to run for the presidency at this moment in history because I believe deeply that we cannot solve the challenges of our time unless we solve them together.
I chose to run for the presidency at this moment in history because I believe deeply that we cannot solve the challenges of our time unless we solve them together.

Source: https://my.barackobama.com/page/content/hisownwords/
I chose to run for the presidency at this moment in history because I believe deeply that we cannot solve the challenges of our time unless we solve them together.

Source: https://my.barackobama.com/page/content/hisownwords/
I chose to run for the presidency at this moment in history because I believe deeply that we cannot solve the challenges of our time unless we solve them together.
I tender my heartfelt gratitude to all of them, while taking full responsibility for all errors . . .

[Mel’čuk1998, 23]
“In all kinds of texts, collocations are indispensable elements with which our utterances are very largely made”

[Kjellmer1987, 10]
Collocations

“Collocation is the way words combine in a language to produce natural-sounding speech and writing”

[Lea and Runcie2002, vii]
“Knowledge that will account for speakers‘ ability to construct and understand phrases and expressions in their language which are not covered by the grammar, the lexicon, and the principles of compositional semantics”

[Fillmore et al.1988, 504]
“Advanced learners of second language have great difficulty with nativelike collocation and idiomaticity. Many grammatical sentences generated by language learners sound unnatural and foreign.”

[Ellis2008]
More examples

- EN open air
- FR plein air ‘full’
- RO aer liber ‘free’
More examples

- EN open air
- FR plein air ‘full’
- RO aer liber ‘free’
More examples

- EN *open air*
- FR *plein air* ‘full’
- RO *aer liber* ‘free’
More examples

- **EN** *ask a question*

- **IT** *fare una domanda*, **ES** *hacer una pregunta*
  ‘make’

- **RO** *a pune o întrebare*, **FR** *poser une question*
  ‘put’
More examples

- EN ask a question
- IT fare una domanda, ES hacer una pregunta
  ‘make’
- RO a pune o intrebari, FR poser une question
  ‘put’
More examples

- EN ask a question
- IT fare una domanda, ES hacer una pregunta
  ‘make’
- RO a pune o intrebari, FR poser une question
  ‘put’
More examples

- EN error occurred
- FR erreur s’est produite
  ‘produced itself’
More examples

- EN error occurred
- FR erreur s’est produite
  ‘produced itself’
EN cheat death

FR frôler la mort
‘brush’
More examples

- EN *cheat death*
- FR *frôler la mort*
  
  ‘brush’
More examples

• EN *reach an agreement*

• FR *parvenir à un accord*
  ‘arrive, get to’

• IT *trovare un accordo*
  ‘find’
More examples

- **EN** reach an agreement
- **FR** parvenir à un accord
  ‘arrive, get to’
- **IT** trovare un accordo
  ‘find’
More examples

- **EN** reach an agreement
- **FR** parvenir à un accord
  ‘arrive, get to’
- **IT** trovare un accordo
  ‘find’
• EN  *money laundering*

• FR  *blanchiment d’argent*
  ‘whitening’

• IT  *lavaggio di denaro*
  ‘washing’
More examples

- EN money laundering
- FR blanchiment d’argent
  ‘whitening’
- IT lavaggio di denaro
  ‘washing’
More examples

- EN money laundering
- FR blanchiment d’argent
  ‘whitening’
- IT lavaggio di denaro
  ‘washing’
Even more examples

- EN narrow majority
- FR courte majorité
  ‘short’
Even more examples

- **EN** narrow majority
- **FR** courte majorité
  ‘short’
- **EN** bring to justice
- **FR** traduire en justice
  ‘translate’
- **FR** urmari în justiție
  ‘follow track’
Even more examples

- EN narrow majority
- FR courte majorité
  ‘short’
- EN bring to justice
- FR traduire en justice
  ‘translate’
- FR urmari în justiție
  ‘follow track’
- EN story breaks
- IN sharp contrast
- strike a deal
- draw criticism
- entertain hope
- experience difficulty
- foot the bill
- meet requirement
- fine weather
- deep impression
- serious injury
Even more examples

courte majorité/narrow majority
etroite collaboration/close cooperation
faible densité/low density
ferme conviction/strong belief
forte augmentation/substantial increase
forte concentration/high concentration
forte pression/heavy pressure
grande attention/great attention
grande diversité/wide range
grande vitesse/high speed
grave erreur/bad error
grossière erreur/great mistake
jeune âge/early age

Automatically extracted collocation equivalents [Seretan and Wehrli2007]
“Collocations make up the lion’s share of the phraseme inventory, and thus they deserve our special attention.”

[Mel’čuk1998, 24]
Importance for Machine Translation

“collocations are the key to producing more acceptable output”

[Orliac and Dillinger 2003, 292]
“collocations are not only considered useful, but also a problem”

[Heylen et al. 1994, 1240]
Collocations act as lexical disambiguators

**to break a record**

- `break (V)` - about 50 senses
- `record (N)` - about 10 senses
Importance for Language Analysis

1. Collocations act as lexical disambiguators

**to break a record**

*break* (V) - about 50 senses  
*record* (N) - about 10 senses  

about 500 potential interpretations

---

[1] Yarowsky1993, 266
Collocations act as lexical disambiguators

"a polysemous word exhibits essentially only one sense per collocation"

[Yarowsky1993, 266]
Collocations act as structural disambiguators.
Collocations act as structural disambiguators

The number of potential parses is exponential in sentence length.
Collocations act as structural disambiguators

The number of potential parses is exponential in sentence length. Collocational knowledge guides the syntactic parsing (e.g., [Wehrli et al.2010]).
Importance for Corpus Linguistics

grimmest traffic jam seems better than man in a traffic jam who curses all transport. a traffic jam in swansea. in enormous traffic jam, and it looked stuck in a traffic jam with your pulse stuck in a traffic jam or being promo. ise every traffic jam will gradually stuck in a traffic jam with their en tting in a traffic jam with such a car iting in a traffic jam. that driver alter the traffic jam in the slightest.

Source: http://www.linguistik-online.de/31_07/danielsson.html
Importance for Lexicography

If you **pay attention** to someone, you watch them, listen to them, or take notice of them. If you **pay no attention** to someone, you behave as if you are not aware of them or as if they are not important.

*pay attention* phrase *V inflects*

More than ever before, the food industry is paying attention to young consumers... Other people walk along the beach at night, so I didn't pay any attention at first...

Source: Collins COBUILD online
NLP applications

- machine translation
- syntactic parsing
- natural language generation
- word sense disambiguation
- topic segmentation
- text summarization
- text classification
- information retrieval
- OCR, speech recognition
1. On collocations

2. Importance of collocations

3. Distinguishing features

4. The need for morphosyntactic analysis

5. Syntax-based extractors

6. Method, results, evaluation

7. Conclusion
Distiguishing features

1. Collocations are partly compositional
Collocations are morphosyntactically flexible

- To buy a book
- To spill the beans (idiom)
- By and large (compound)
**make proposal**

A proposal for the financing of the variable costs will be made to the Committee . . .

**submit proposal**

A joint proposal which addressed such elements as notification, consultations, conciliation and mediation, arbitration, panel procedures, technical assistance, adoption of panel reports and GATTs surveillance of their implementation was submitted on behalf of fourteen participants.
Syntactic flexibility

**make proposal**

A *proposal* for the financing of the variable costs will be *made* to the Committee . . .

**submit proposal**

A joint *proposal* which addressed such elements as notification, consultations, conciliation and mediation, arbitration, panel procedures, technical assistance, adoption of panel reports and GATTs surveillance of their implementation was *submitted* on behalf of fourteen participants.
Syntactic flexibility

**serious problem**

*We all need to renounce the use of arms in order to be able to address the country’s serious political, social and economic problems.*

**important issue**

*The issue of new technologies and their application in education naturally generates considerable interest and is extremely important.*
Syntactic flexibility

**serious problem**

We all need to renounce the use of arms in order to be able to address the country’s serious political, social and economic problems . . .

**important issue**

The issue of new technologies and their application in education naturally generates considerable interest and is extremely important.
Notant en outre que le paragraphe 5 de l’Acte final reprenant les résultats des Négociations commerciales multilatérales du Cycle d’Uruguay (ci-après dénommés respectivement l’“Acte final” et le “Cycle d’Uruguay”) dispose que . . .
Identification approaches

1. Approaches based on linear proximity
   Definition:
   
   “Collocation is the cooccurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening.” [Sinclair1991, 170].

2. Approaches based on structural proximity
   Definition:
   
   “lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other” [Bartsch2004, 76]
Identification approaches

1. Approaches based on linear proximity

Definition:

“*Collocation is the cooccurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening.*” [Sinclair1991, 170].

2. Approaches based on structural proximity

Definition:

“*lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in a direct syntactic relation with each other*” [Bartsch2004, 76]
1. On collocations

2. Importance of collocations

3. Distinguishing features

4. The need for morphosyntactic analysis

5. Syntax-based extractors

6. Method, results, evaluation

7. Conclusion
Inflection

1 type: *jouer - rôle*

50 forms: *jouent rôle, rôle joué, joue rôles, ...*
Lexical ambiguity

**human**
- Noun
- Adjective

**rights**
- Noun
- Adjective
- Verb
- Adverb

**human – rights**
- Noun – Noun
- Noun – Verb
- Adjective – Noun

...
Inversion

set – record

*Michael Phelps* sets *all-time Olympic record.*

*A new world record has been set.*
Inversion
Long-distance dependencies

donner – exemple ‘give – example’

Le visionnaire a donné, lors de sa conférence magistrale, à l’occasion de la remise du Prix Latsis 2011 aux différents lauréats, le mercredi 30 novembre 2011, dans la salle Piaget de l’Université de Genève à Uni-Dufour, devenu trop petite pour accueillir le monde scientifique et le public venus de tous les coins et recoins de la Suisse, l’exemple du professeur et ancien président sénégalais le poète Léopold Sédar Senghor qui maîtrisait autant la culture du pays de Marianne que la langue de Molière avec perfection pour devenir le premier Noir membre de l’Académie française.
Long-distance dependencies

put – off

It is still wise to find a tactful way of putting those who have a virus or the flu, or any other problem off for a little while longer.
question - ask

Any **question asked** during the selection and interview process must be related to the job and the performance of that job.

The **question asked** if the grant funding could be used as start-up capital to develop this project.
structural ambiguity

**question - ask**

Any **question asked** during the selection and interview process must be related to the job and the performance of that job.

*The question asked* if the grant funding could be used as start-up capital to develop this project.
Typical extraction method

**Sliding window (Baseline)**

1. Take into account all possible combinations within a 5-word collocational span
2. Apply an *association measure* to filter out noise and retain best candidates on top of the output list
Typical extraction method

**Sliding window (Baseline)**

1. Take into account all possible combinations within a 5-word collocational span
2. Apply an *association measure* to filter out noise and retain best candidates on top of the output list

**WARNING**

MANUALLY VALIDATE RESULTS BEFORE USE
The multilingual challenge

In English, many syntactic relations are realised in a 5-word window.

But what about freer word order languages?
The multilingual challenge

German

“Some properties of the German language make the task of extracting V-N collocations from German text corpora more difficult than for English corpora.”
[Breidt1993, 77]

“the assumption that a “semantic agent [...] is principally used before the verb” and a “semantic object [...] is used after it” as described in Smadja (1991a:180) does not hold for German. Therefore, complicated parsing is necessary to distinguish subject-verb from object-verb combinations.”
[Breidt1993, 77]
The multilingual challenge

German

“Some properties of the German language make the task of extracting V-N collocations from German text corpora more difficult than for English corpora.”
[Breidt1993, 77]

“the assumption that a “semantic agent [...] is principally used before the verb” and a “semantic object [...] is used after it” as described in Smadja (1991a:180) does not hold for German. Therefore, complicated parsing is necessary to distinguish subject-verb from object-verb combinations.”
[Breidt1993, 77]
The multilingual challenge

Korean

“The free order of Korean makes it hard to identify collocations.”
[Kim et al. 1999, 71]

“Unfortunately, the approach for English has several limitations to work on Korean structure”
[Kim et al. 1999, 71]
The multilingual challenge

Korean

“The free order of Korean makes it hard to identify collocations.”
[Kim et al. 1999, 71]

“Unfortunately, the approach for English has several limitations to work on Korean structure”
[Kim et al. 1999, 71]
“Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure.

[Smadja1993, 151]
Solution

“Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure. However, in practice, free-style texts contain a great deal of nonstandard features over which automatic parsers would fail.

[Smadja1993, 151]
“Ideally, in order to identify lexical relations in a corpus one would need to first parse it to verify that the words are used in a single phrase structure. However, in practice, free-style texts contain a great deal of nonstandard features over which automatic parsers would fail. [...] This fact is being seriously challenged by current research [...] and might not be true in the near future.”
[Smadja1993, 151]
“with recent significant increases in parsing efficiency and accuracy, there is no reason why explicit parse information should not be used” [Pearce2002, 1530]
“Ideally, a full syntactic analysis of the source corpus would allow us to extract the cooccurrence directly from parse trees” [Evert2004, 31]
1. On collocations
2. Importance of collocations
3. Distinguishing features
4. The need for morphosyntactic analysis
5. Syntax-based extractors
6. Method, results, evaluation
7. Conclusion
<table>
<thead>
<tr>
<th>Author and Year</th>
<th>Language(s)</th>
<th>Type</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lin (1998, 1999)</strong> – English</td>
<td></td>
<td>dependency parser (unspecified)</td>
<td>sentences shorter than 25 words; 9.7% errors in top results; 6 types: N-D, N-A, N-N, V-N, N-V, V-Adv</td>
</tr>
<tr>
<td><strong>Wu and Zhou (2003); Lü and Zhou (2004)</strong> – English, Chinese</td>
<td></td>
<td>syntactic parser (NLPWin, Microsoft Research)</td>
<td>7.85% errors in top results; 3 types: V-O, N-A, V-Adv</td>
</tr>
<tr>
<td><strong>Villada Moirón (2005)</strong> – Dutch</td>
<td></td>
<td>dependency parser (Alpino)</td>
<td>sentences shorter than 20 words; many PP-attachment errors; parser only used for chunking; 2 types: P-N-P and PP-V</td>
</tr>
<tr>
<td><strong>Orliac and Dillinger (2003)</strong> – English</td>
<td></td>
<td>deep parser (Logos)</td>
<td>limited grammatical coverage (does not handle relatives); 3 types: S-V, V-O, V-P-N</td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td></td>
<td></td>
<td>Zinsmeister and Heid (2003), Schulte im Walde (2003) – German, statistical parser (LoPar); Charest et al. (2007) – French, dependency parser (Antidote)</td>
</tr>
</tbody>
</table>
Lin (1998, 1999) – English
- dependency parser (unspecified)
- sentences shorter than 25 words
- 9.7% errors in top results
- 6 types: N-D, N-A, N-N, V-N, N-V, V-Adv

- syntactic parser (NLPWin, Microsoft Research)
- 7.85% errors in top results
- 3 types: V-O, N-A, V-Adv

Villada Moirón (2005) – Dutch
- dependency parser (Alpino)
- sentences shorter than 20 words
- many PP-attachment errors; parser only used for chunking
- 2 types: P-N-P and PP-V

Orliac and Dillinger (2003) – English
- deep parser (Logos)
- limited grammatical coverage (does not handle relatives)
- 3 types: S-V, V-O, V-P-N

Others
Zinsmeister and Heid (2003), Schulte im Walde (2003) – German, statistical parser (LoPar)
Charest et al. (2007) – French, dependency parser (Antidote)
Lin (1998, 1999) – English
- dependency parser (unspecified)
- sentences shorter than 25 words
- 9.7% errors in top results
- 6 types: N-D, N-A, N-N, V-N, N-V, V-Adv

- syntactic parser (NLPWin, Microsoft Research)
- 7.85% errors in top results
- 3 types: V-O, N-A, V-Adv

Villada Moirón (2005) – Dutch
- dependency parser (Alpino)
- sentences shorter than 20 words
- many PP-attachment errors; parser only used for chunking
- 2 types: P-N-P and PP-V

Orliac and Dillinger (2003) – English
- deep parser (Logos)
- limited grammatical coverage (does not handle relatives)
- 3 types: S-V, V-O, V-P-N

Others
- Zinsmeister and Heid (2003), Schulte im Walde (2003) – German, statistical parser (LoPar)
- Charest et al. (2007) – French, dependency parser (Antidote)
<table>
<thead>
<tr>
<th>Study</th>
<th>Language(s)</th>
<th>Type</th>
<th>System</th>
<th>Accuracy/Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Villada Moirón (2005)</td>
<td>Dutch</td>
<td>dependency parser</td>
<td>Alpino</td>
<td>sentences shorter than 20 words, many PP-attachment errors, parser only used for chunking, 2 types: P-N-P and PP-V</td>
</tr>
<tr>
<td>Orliac and Dillinger (2003)</td>
<td>English</td>
<td>deep parser</td>
<td>Logos</td>
<td>limited grammatical coverage, (does not handle relatives), 3 types: S-V, V-O, V-P-N</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zinsmeister and Heid (2003), Schulte im Walde (2003)</td>
<td>German, statistical parser (LoPar)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charest et al. (2007)</td>
<td>French</td>
<td>dependency parser</td>
<td>Antidote</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Language(s)</td>
<td>Type</td>
<td>Specific Details</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------</td>
<td>-----------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Lin (1998, 1999)</td>
<td>English</td>
<td>dependency parser</td>
<td>unspecified</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>sentences shorter than 25 words</td>
<td>9.7% errors in top results</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 types: N-D, N-A, N-N, V-N, N-V, V-Adv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu and Zhou (2003); Lü and Zhou (2004)</td>
<td>English, Chinese</td>
<td>syntactic parser</td>
<td>NLPWin, Microsoft Research</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.85% errors in top results</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 types: V-O, N-A, V-Adv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Villada Moirón (2005)</td>
<td>Dutch</td>
<td>dependency parser</td>
<td>Alpino</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>sentences shorter than 20 words</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>many PP-attachment errors; parser only used for chunking</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 types: P-N-P and PP-V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orliac and Dillinger (2003)</td>
<td>English</td>
<td>deep parser</td>
<td>Logos</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>limited grammatical coverage (does not handle relatives)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 types: S-V, V-O, V-P-N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td>Zinsmeister and Heid (2003), Schulte im Walde (2003) – German, statistical parser (LoPar)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Charest et al. (2007) – French, dependency parser (Antidote)</td>
<td></td>
</tr>
</tbody>
</table>
The LATL syntax-based collocation extractor: FipsCo

Goldman et al. (2001) – English, French

- deep parser (Fips)
- broad grammatical coverage
- many types of collocation configurations

- FipsCo precedes many syntax-based extractors, and overcomes limitations pertaining to parsing robustness, precision, coverage, as well as limitations regarding the list of supported syntactic types.

- It has been developed mainly as a CAT tool for WTO translators in the project “Linguistic Analysis and Collocation Extraction”.

- Initially available for English and French, it was further extended to Spanish, Italian, Greek, German and Romanian.
The LATL syntax-based collocation extractor: FipsCo

Goldman et al. (2001) – English, French

- deep parser (Fips)
- broad grammatical coverage
- many types of collocation configurations

FipsCo precedes many syntax-based extractors, and overcomes limitations pertaining to parsing robustness, precision, coverage, as well as limitations regarding the list of supported syntactic types.

It has been developed mainly as a CAT tool for WTO translators in the project “Linguistic Analysis and Collocation Extraction”.

Initially available for English and French, it was further extended to Spanish, Italian, Greek, German and Romanian.
The LATL syntax-based collocation extractor: FipsCo

Goldman et al. (2001) – English, French

- deep parser (Fips)
- broad grammatical coverage
- many types of collocation configurations

- FipsCo precedes many syntax-based extractors, and overcomes limitations pertaining to parsing robustness, precision, coverage, as well as limitations regarding the list of supported syntactic types.

- It has been developed mainly as a CAT tool for WTO translators in the project “Linguistic Analysis and Collocation Extraction”.

- Initially available for English and French, it was further extended to Spanish, Italian, Greek, German and Romanian.
Goldman et al. (2001) – English, French

- deep parser (Fips)
- broad grammatical coverage
- many types of collocation configurations

FipsCo precedes many syntax-based extractors, and overcomes limitations pertaining to parsing robustness, precision, coverage, as well as limitations regarding the list of supported syntactic types.

- It has been developed mainly as a CAT tool for WTO translators in the project “Linguistic Analysis and Collocation Extraction”.
- Initially available for English and French, it was further extended to Spanish, Italian, Greek, German and Romanian.
It is very hard in these difficult circumstances for Georgia to observe all its current human rights obligations. That is why we are calling on the Commission to act here. We have made requests and have largely reached cross-party agreement in Parliament that something should be done. We need to work together to come up with solutions which will promote a peaceful way forward.

Mr President, I would like to firmly support Mrs Schleicher, who has worked with great commitment on behalf of the Caucasian region and in this case Georgia in particular. We all know what these reprisals are about. On the one hand they are linked with geo-strategic interests and with energy and oil interests, and on the other hand they are associated with

asumido en el Consejo de Europa.

En estas difíciles circunstancias, resulta difícil para Georgia cumplir todos sus deberes actuales. Por esta razón, pedimos a la Comisión que haga algo en este punto. Hemos planteado exigencias y en el Parlamento estamos también de acuerdo, por encima de todos los Grupos, en que es preciso que se haga algo. Busquemos conjuntamente soluciones que posibiliten salidas pacíficas.

Señor Presidente, quisiera apoyar expresamente a nuestra colega, la Sra. Schleicher, que está trabajando con un gran compromiso en pro de la región del Cáucaso y, en este caso, de una manera especial, en pro de Georgia. Todos sabemos con qué tienen que ver estas

Details
Type: Verb-Object
Score: 97.34
Rank: 1
Status: New collocation!
reach agreement

build a democratic society and seek a peaceful future, and to support Georgia in observing human rights and complying with the other obligations it has entered into within the Council of Europe.

It is very hard in these difficult circumstances for Georgia to observe all its current human rights obligations. That is why we are calling on the Commission to act here. We have made requests and have largely reached cross-party agreement in Parliament that something should be done. We need to work together to come up with solutions which will promote a peaceful way forward.

Mr President, I would like to firmly support Mrs Schleicher, who has worked with great commitment on behalf of the Caucasian region and in this case Georgia in particular. We all know what these reprisals are about. On the one hand they are linked with geo-strategic interests and with energy and oil interests, and on the other hand they are associated with the brutal war being waged to eradicate the people of Chechnya. It is very clear to us what is being attempted here - to destabilise Georgia, just as Chechnya has been destabilised. We all have an interest in preserving what is in any case very fragile nationhood and in supporting the first faltering steps towards democracy in this region and in Georgia.

That is why I would like to call on the Commissioner to pay even greater attention to...
Filter Collocations

Collocation Filter

Key1:  

Key2:  agreement

Type:
- Adjective-Noun
- Noun-Adjective
- Noun-Noun(head)
- Subject-Verb
- Verb-Object
- Verb-Prep
- Verb-Prep-Arg
- Noun-Pred-Adjective
- Compound
- Noun-Prep-Noun
- Verb-Adjective
- Verb-Adverb
- Adverb-Adjective
- Adjective-Prep

Display
- new collocations
- validated collocations only
- collocations in lexic
- mark validated collocations

Filter
- score: > 0.0 to 0.0 (undef.)
- frequency: 1 to 0 (undef.)
- range: between 1 and 500
- considering: tokens
- types
- order: score

Apply  Cancel
Validate Collocations

Details
Collocation: reach agreement
Type: Verb-Object
Key1: reach
Key2: agreement

Sample context:
At Nice, 15 Member States, each - understandably - focussing on its own national interests, were able to reach only an imperfect agreement.

Translation: trovare un accordo
Language: Italian

Translation context:
A Nizza quindici Stati membri, ciascuno dei quali concentrato, anche comprensibilmente, sui propri interessi nazionali, non hanno potuto trovare che un accordo imperfetto.
Display Collocations

Context in file: clean_challenges_1000.odc


- "Our country faces profound challenges that must be addressed in order to ensure .... growing challenges to our security. The volume of data is staggering. ..." (http://nvac.pnl.gov/docs/RD_Agenda_NVAC_chapter1.pdf)

- "Challenges are sent by the NS from time to time, the first one shortly after the initial status has been set with CHG. There’s no regular period to the ..." (http://www.hypothetic.org/docs/msn/notification/pings_challenges.php)

- "explores some of the challenges that organizations must overcome to be .... management challenges, particularly when the focus is security. ..." (http://www.cert.org/archive/pdf/ESMchallenges.pdf)

- "Apr 10, 2008 ... Phoenix offers Martinsville-like challenges. ... Phoenix also features challenging doglegs and a unique challenge in the corners, ..." (http://msn.foxsports.com/nascar/story/8003374/Phoenix-offers-Martinsville-like-challenges)
Collocation extraction

Language: English
Association measure (AM): Log Likelihood Ratio
AM score (min.): 0.0
Occurrences (min.): 3

Input file:
Choose File: No file chosen
File encoding: ANSI, UTF-8. Size limit: 500 000 words. The extraction method is described here.

[Optional] Enter an e-mail address if you wish to be notified when processing is completed:

Extract

Results:
FipsCoView

English

Word: challenge

Occurrences: min: 3

Score: min: 0.0

Results:

→ Noun. 16 collocations
Adjective-Noun
new ~ 12x great ~ 10x major ~ 7x enormous ~ 5x demographic ~ 5x huge ~ 4x legal ~ 3x

Subject-Verb
~ to be 9x ~ to be about 4x

Verb-Object
to be ~ 13x to meet ~ 12x to tackle ~ 6x to address ~ 5x to pose ~ 3x to accept ~ 3x

Noun-Prep-Noun
~ of globalization 4x
Experiments

Languages

English, French, Italian, Spanish [Seretan and Wehrli2009], German, Greek [Michou and Seretan2009], Romanian [Seretan and Wehrli2010]

Corpora

WTO translation archives, The Economist, Canadian Hansard, Le Monde, Europarl, Mesagerul, WWW (Web as a corpus) ...
Display Collocations

Collocations (100)

Source: D:\corpus\ro\jurnalux162.odc

Avea loc
unione europeană
stat unit
nație unită
stat unit
altă parte
atrage atenție
punct de vedere
scurt timp
imigrant ilegal
mia de metru
fi vorbă
miliard de dolar
milion de dolar
campionat mondial
datorie istorică
alegere preșidențială
da în judecată
casă albă
cupă mondială
persoană cu handicap
săptămână trecută
campion mondial
test nuclear

Romanian

Crt: 1 of 52

Details
Type: Verb - Object
Score: 339.90
Rank: 1

Options ...
Validate ...
Cancel
Exploitation

- Lexical acquisition
  - for the Fips parser
  - for the Its-2 MT system [Wehrli et al.2009]

- Text Summarization [Seretan2011]

- Syntactic Parsing [Seretan and Wehrli2011]

- Terminology assistance [Wehrli2003, Wehrli2006]

- Teaching

- Research
Lexical acquisition

Collocation: weapon of mass destruction

Collocation information:
- Index: 141005079 - 'weapon of mass dest
- Lexemes: 111041841, 141000293
- Type: nom prép nom
- Prep: of
- Features: [detLessComp]

Selection of lexemes:
- Lexeme 1: weapon : N [111041841] -
- Lexeme 2: mass destruction : [141000293] -
Lexical acquisition

Collocations

Collocation: weapon of mass destruction

English

Search

Unsafe

Collocation information

Index: 141005079 - 'weapon of mass dest

Lexemes:
111041841
141000293

Type: nom prép nom

Prep: of

Features: [detLessCompl]

Frequency: 0

Selection of lexemes

Lexeme 1

weapon: N [111041841]-

Info

Alternatives

Feature Selector

detLessCompl

pluralColloc
detLessCompl
detColloc
notDetLessCompl
pluralColloc
possessiveCompl
bareNounCompl
frozenCompl
nounPossNoun
indefiniteDetCompl
compound

Remove

Update / Insert

Cancel

OK
À la recherche du mot juste

Découvrez, à l'aide de l'outil *FipsCoView* développé au LATL, des combinaisons typiques de mots, telles qu'elles ont été extraites à partir de textes multilingues grâce à une méthode automatique qui allie analyse syntaxique et calcul statistique.

Accéder

© LATL 2011 | Contact

Retour aux Activités et ressources pédagogiques UNIGE
Terminology assistance

The leading software on the market for money laundering research - Mozilla Firefox

Extranet

ISO 9001:2008

CERTIFIED MANAGEMENT SYSTEM

efficient and integrated research strategy is now becoming a crucial competitive factor. As well as compliance with legal provisions, other determining factors include flexibility and low cost-and-time overhead for implementing and adjusting the research strategy. This enables companies to effectively counter new developments in money laundering.

Risk-based prevention of money laundering

This flexibility and efficiency is provided by SironAML, the integrated anti-money laundering solution. The starting point for ongoing adjustment of the research processes is a reassessment of customer- and product-specific money laundering risks within the context of the enterprise-wide threat analysis. Preventative measures derived from this (IT research rules) can be independently created or modified by the money laundering/compliance officers in an intuitive rule editor in SironAML. Predefined standard research rules for compliance with FATF (Financial Action Task Force) recommendations, among others, are already integrated in the software.

Customer-oriented Case Management

Customer and transaction data is analysed in SironAML by means of individually definable research rules. All rule infringements are filtered out and forwarded to the integrated Case Management for further analysis by the money laundering/compliance officer. All suspicious transactions
Fips parser [Wehrli2007] – Sample output
Method

Fips parser [Wehrli2007] – Sample output
Fips - Key facts

**Constituent**

- simplified X-bar structure \([x_p L X R]\) (no intermediate level)
- X – lexical head (A, N, V, D, P, Conj, ...)
- L/R – lists of left/right subconstituents

**Manually-built lexica**

- detailed morphosyntactic and semantic information: selectional properties, subcategorization information, syntactico-semantic features likely to influence the syntactic analysis

**Algorithm - main operations**

- *Project*: assignment of constituent structures to lexical entries
- *Merge*: combination of adjacent constituents
- *Move*: creation of chains by linking surface positions of “moved” constituents to their corresponding canonical positions.
Fips - Key facts

Constituent

- simplified X-bar structure \([X_P \ L \ X \ R]\) (no intermediate level)
  - \(X\) – lexical head (A, N, V, D, P, Conj, ...)
  - L/R – lists of left/right subconstituents

Manually-built lexica

- detailed morphosyntactic and semantic information: selectional properties, subcategorization information, syntactico-semantic features likely to influence the syntactic analysis

Algorithm - main operations

- **Project**: assignment of constituent structures to lexical entries
- **Merge**: combination of adjacent constituents
- **Move**: creation of chains by linking surface positions of “moved” constituents to their corresponding canonical positions.
## Fips - Key facts

### Constituent
- Simplified X-bar structure $[x_P \ L \ X \ R]$ (no intermediate level)
  - $X$ – lexical head (A, N, V, D, P, Conj, ...)
  - $L/R$ – lists of left/right subconstituents

### Manually-built lexica
- Detailed morphosyntactic and semantic information: selectional properties, subcategorization information, syntactico-semantic features likely to influence the syntactic analysis

### Algorithm - main operations
- **Project**: assignment of constituent structures to lexical entries
- **Merge**: combination of adjacent constituents
- **Move**: creation of chains by linking surface positions of “moved” constituents to their corresponding canonical positions.
Method – Stage 1: Candidate selection

![Diagram of candidate selection method]

The diagram illustrates the candidate selection stage in collocation extraction. It shows a tree structure with nodes labeled as XP, X, L, X_j, and R, indicating the selection criteria for candidate collocations.
Method – Stage 1: Candidate selection
Method – Stage 1: Candidate selection

[Diagram showing a tree structure with nodes and arrows indicating the selection process.]
Method – Stage 1: Candidate selection

1. **Lexical filter:**
   rule out auxiliary and modal verbs, proper nouns, common nouns representing titles (*Mr.*)

2. **Structural filter:**
   predicate-argument relation in the arguments table of predicates combinations \(<X, \text{head of item in L/R}>\) in a given syntactic relation, e.g., head-modifier, noun-adjective in FP (functional phrase)
Method – Stage 1: Candidate selection

1. **Lexical filter:**
   rule out auxiliary and modal verbs, proper nouns, common nouns representing titles (Mr.)

2. **Structural filter:**
   predicate-argument relation in the arguments table of predicates combinations \(<X, \text{head of item in L/R}>\) in a given syntactic relation, e.g., head-modifier, noun-adjective in FP (functional phrase)
Method – Stage 1: Candidate selection

Syntactic patterns

- adjective-noun
  - heavy smoker
- noun-[predicate]-adjective
  - effort [be] devoted
- noun-noun
  - suicide attack
- noun-preposition-noun
  - round of negotiations
- noun-preposition
  - inquiry into
- adjective-preposition
  - crazy about
- subject-verb
  - war breaks
- verb-object
  - meet requirement
- verb-preposition-argument
  - bring to boil
- verb-preposition
  - point out
- adverb-verb
  - fully support
- adverb-adjective
  - highly important
Method – Stage 2: Candidate ranking

<table>
<thead>
<tr>
<th>AM</th>
<th>Explicit formula</th>
<th>AM</th>
<th>Explicit formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>( \frac{(a + b + c + d)(ad - bc)^2}{(a + b)(a + c)(b + d)(c + d)} )</td>
<td>Odds ratio</td>
<td>( \log \frac{ad}{bc} )</td>
</tr>
<tr>
<td>Dice coefficient</td>
<td>( \frac{2a}{2a + b + c} )</td>
<td>Poisson-Stirling</td>
<td>( a \left( \log a - \log \frac{(a + b)(a + c)}{(a + b + c + d)} - 1 \right) )</td>
</tr>
<tr>
<td>Geometric mean</td>
<td>( \frac{a}{\sqrt{(a + b)(a + c)}} )</td>
<td>Relative frequency</td>
<td>( \frac{a}{a + b + c + d} )</td>
</tr>
<tr>
<td>Jaccard</td>
<td>( \frac{a}{a + b + c} )</td>
<td>Relative risk</td>
<td>( \log \frac{a(b + d)}{b(a + c)} )</td>
</tr>
<tr>
<td>Lindell</td>
<td>( \frac{ad - bc}{(a + c)(b + d)} )</td>
<td>Salience</td>
<td>( \log_2 \frac{a(a + b + c + d)}{(a + b)(a + c)} \log_2 a )</td>
</tr>
<tr>
<td>Log-likelihood ratios (LLR)</td>
<td>( 2(a \log a + b \log b + c \log c + d \log d - \frac{a + b}{(a + b) \log(a + b) - (a + c) \log(a + c) - \frac{b + d}{(b + d) \log(b + d) - (c + d) \log(c + d) + \frac{a + b + c + d}{(a + b + c + d) \log(a + b + c + d)}} )</td>
<td>t-score</td>
<td>( \frac{a(a + b + c + d) - (a + b)(a + c)}{(a + b + c + d) \sqrt{a}} )</td>
</tr>
<tr>
<td>Mutual information (MI)</td>
<td>( \log_2 \frac{a(a + b + c + d)}{(a + b)(a + c)} )</td>
<td>z-score</td>
<td>( \frac{a(a + b + c + d) - (a + b)(a + c)}{(a + b + c + d) \sqrt{(a + b)(a + c)}} )</td>
</tr>
</tbody>
</table>
Method – Summing up:

Baseline

1. Candidate identification:
   Take into account all possible combinations within a 5-word collocational span

2. Candidate ranking:
   Apply an association measure to filter out noise and retain best candidates on top of the output list

Syntax-based extraction

1. Candidate identification:
   Take into account syntactically bound combinations, according to the parse tree built by Fips for the input sentence

2. Candidate ranking:
   Apply an association measure to retain best candidates on top of the output list
Method – Summing up:

Baseline

1. Candidate identification:
   Take into account **all** possible combinations within a 5-word collocational span

2. Candidate ranking:
   Apply an association measure to **filter out noise and** retain best candidates on top of the output list

Syntax-based extraction

1. Candidate identification:
   Take into account **syntactically bound** combinations, according to the parse tree built by Fips for the input sentence

2. Candidate ranking:
   Apply an association measure to retain best candidates on top of the output list
a very simple question which everyone in this country would like to ask

No doubt that will be partly due to the great contribution you and your colleagues in the Chair will make.

The provincial government made a very difficult but well balanced decision that enhances environmental, economic and social values for the area.

It is the new government’s responsibility to tackle with conviction and fairness the complex problems facing Canada ...

at a cost of $5 billion that is chiefly being met by South Korea and Japan
Results – Tokens identified

*a very simple* question which everyone in this country would like to ask

No doubt that will be partly due to the great contribution you and your colleagues in the Chair will make.

The provincial government made a very difficult but well balanced decision that enhances environmental, economic and social values for the area.

It is the new government’s responsibility to tackle with conviction and fairness the complex problems facing Canada ...

at a cost of $5 billion that is chiefly being met by South Korea and Japan
a very simple question which everyone in this country would like to ask

No doubt that will be partly due to the great contribution you and your colleagues in the Chair will make.

The provincial government made a very difficult but well balanced decision that enhances environmental, economic and social values for the area.

It is the new government’s responsibility to tackle with conviction and fairness the complex problems facing Canada ...

at a cost of $5 billion that is chiefly being met by South Korea and Japan
a very simple question which everyone in this country would like to ask.

No doubt that will be partly due to the great contribution you and your colleagues in the Chair will make.

The provincial government made a very difficult but well balanced decision that enhances environmental, economic and social values for the area.

It is the new government’s responsibility to tackle with conviction and fairness the complex problems facing Canada ...

at a cost of $5 billion that is chiefly being met by South Korea and Japan.
a very simple question which everyone in this country would like to ask

No doubt that will be partly due to the great contribution you and your colleagues in the Chair will make.

The provincial government made a very difficult but well balanced decision that enhances environmental, economic and social values for the area.

It is the new government’s responsibility to tackle with conviction and fairness the complex problems facing Canada ...

at a cost of $5 billion that is chiefly being met by South Korea and Japan
Results – Syntactic environments

- **passivization:**

  > I see that amendments to the report by Mr Méndez de Vigo and Mr Leinen have been tabled on this subject.

- **relativization:**

  > The communication devotes no attention to the impact the newly announced policy measures will have on the candidate countries.

- **interrogation:**

  > What impact do you expect this to have on reducing our deficit and our level of imports?

- **cleft constructions:**

  > It is a very pressing issue that Mr Sacrédeus is addressing.

- **coordinated clauses:**

  > This motion implies that somehow the current income tax laws on alimony and maintenance payments are unfair, contribute to the problem and therefore should be amended.
Results – Syntactic environments

- passivization:

  *I see that amendments to the report by Mr Méndez de Vigo and Mr Leinen have been tabled on this subject.*

- relativization:

  *The communication devotes no attention to the impact the newly announced policy measures will have on the candidate countries.*

- interrogation:

  *What impact do you expect this to have on reducing our deficit and our level of imports?*

- cleft constructions:

  *It is a very pressing issue that Mr Sacrédeus is addressing.*

- coordinated clauses:

  *This motion implies that somehow the current income tax laws on alimony and maintenance payments are unfair, contribute to the problem and therefore should be amended.*
passivization:

*I see that amendments to the report by Mr Méndez de Vigo and Mr Leinen have been tabled on this subject.*

relativization:

*The communication devotes no attention to the impact the newly announced policy measures will have on the candidate countries.*

interrogation:

*What impact do you expect this to have on reducing our deficit and our level of imports?*

cleft constructions:

*It is a very pressing issue that Mr Sacrédeus is addressing.*

coordinated clauses:

*This motion implies that somehow the current income tax laws on alimony and maintenance payments are unfair, contribute to the problem and therefore should be amended.*
Results – Syntactic environments

- passivization:
  
  I see that **amendments** to the report by Mr Méndez de Vigo and Mr Leinen have been **tabled** on this subject.

- relativization:
  
  The communication devotes no attention to the **impact** the newly announced policy measures will **have** on the candidate countries.

- interrogation:
  
  What **impact** do you expect this to **have** on reducing our deficit and our level of imports?

- cleft constructions:
  
  It is a very pressing **issue** that Mr Sacrédeus is **addressing**.

- coordinated clauses:
  
  This **motion** implies that somehow the current income tax laws on alimony and maintenance payments are unfair, contribute to the problem and therefore should be **amended**.
passivization:

*I see that amendments to the report by Mr Méndez de Vigo and Mr Leinen have been tabled on this subject.*

relativization:

*The communication devotes no attention to the impact the newly announced policy measures will have on the candidate countries.*

interrogation:

*What impact do you expect this to have on reducing our deficit and our level of imports?*

cleft constructions:

*It is a very pressing issue that Mr Sacrédeus is addressing.*

coordinated clauses:

*This motion implies that somehow the current income tax laws on alimony and maintenance payments are unfair, contribute to the problem and therefore should be amended.*
Evaluation experiment 1: $n$-best evaluation

**Corpus**
- Hansard
  - (Canadian Parliament debates)

**Language**
- French

**Size**
- $\sim 1.2$ M words

**Methods**
- Baseline
- Syntax-based

**Results**
- Significance list

**Evaluation**
- 500 types
- top level
- total: 1000 types
- 3 evaluators/method

**Annotation**
- (-gram) erroneous
- (+gram) (-lex) regular
- (+lex) interesting

Fleiss $\kappa = 0.50$ (Baseline)
Fleiss $\kappa = 0.39$ (Syntax-Based)
(moderate/fair agreement)
Evaluation experiment 1: \( n \)-best evaluation

**Corpus**
- Hansard (Canadian Parliament debates)

**Language**
- French

**Size**
- \( \sim 1.2 \text{ M words} \)

**Methods**
- Baseline
- Syntax-based

**Evaluation**
- 500 types
- top level
- total: 1000 types
- 3 evaluators/method

**Results**

**Annotation**

\[
\begin{array}{|c|c|}
\hline
\text{(-gram)} & \text{erroneous} \\
\text{(+gram)} & \text{(-lex) regular} \\
& \text{(+lex) interesting} \\
\hline
\end{array}
\]

Fleiss \( \kappa = 0.50 \) (Baseline)
Fleiss \( \kappa = 0.39 \) (Syntax-Based)
(moderate/fair agreement)
Evaluation experiment 1: $n$-best evaluation

**Corpus**
- Hansard (Canadian Parliament debates)

**Language**
- French

**Size**
- $\sim 1.2$ M words

**Methods**
- Baseline
- Syntax-based

**Evaluation**
- 500 types top level
- total: 1000 types
- 3 evaluators/method

**Results**
- Significance list

**Annotation**
- (-gram) erroneous
- (+gram) (-lex) regular
- (+lex) interesting

Fleiss $\kappa = 0.50$ (Baseline)
Fleiss $\kappa = 0.39$ (Syntax-Based) (moderate/fair agreement)
## Annotation examples

<table>
<thead>
<tr>
<th>(-gram) erroneous</th>
<th>petites entreprises (*V-O)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+gram) regular</td>
<td>aide à rénovation</td>
</tr>
<tr>
<td>(+lex) interesting</td>
<td>aborder sujet</td>
</tr>
</tbody>
</table>
Confusion examples

attention particulière, bref délai, grave problème, lésion corporelle, offrir service, avoir droit, avoir droit, avoir honneur, créer emploi
(14.0% of the cases)
Confusion examples

**attention particulière, bref délai, grave problème, lésion corporelle, offrir service, avoir droit, avoir droit, avoir honneur, créer emploi**
(14.0% of the cases)

**an prochain, dernier année, écouter discours, fin de semaine, monde entier, offre finale, relation de travail** (19.3% of the cases)
Evaluation results

Outer ring – baseline; inner ring – syntax-based method

Statistical significance: $+\text{gram } t(982) = 10.78, p < 0.001$; $+\text{lex } t(982) = 2.90, p < 0.01$
Evaluation results

Grammatical precision by sets of 50 pairs
Evaluation experiment 2: stratified evaluation

**Corpus**
- Europarl (Koehn, 2005)

**Languages**
- French
- English
- Italian
- Spanish

**Size**
- on average, \( \sim 3.7 \text{ M words/language} \)

**Methods**
- Baseline
- Syntax-based

**Evaluation**
- 50 types
- 5 levels (0-10%)
- total: 2000 types
- 2 evaluators/method

**Results**

**Annotation**

<table>
<thead>
<tr>
<th>(-gram) erroneous</th>
<th>(+gram) (+lex)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-lex) regular</td>
<td>named entity</td>
</tr>
<tr>
<td></td>
<td>compound</td>
</tr>
<tr>
<td></td>
<td>idiom</td>
</tr>
<tr>
<td></td>
<td>collocation</td>
</tr>
</tbody>
</table>

Cohen’s \( \kappa = 0.61 \) (significant agreement)
Evaluation experiment 2: stratified evaluation

**Corpus**
- Europarl
  - (Koehn, 2005)

**Languages**
- French
- English
- Italian
- Spanish

**Size**
- on average, \(\sim 3.7\) M words/language

**Evaluation**
- 50 types
- 5 levels (0-10%)
- total: 2000 types
- 2 evaluators/method

**Methods**
- Baseline
- Syntax-based

**Results**
- Significance list

**Annotation**
<table>
<thead>
<tr>
<th>(-gram) erroneous</th>
<th>(+gram) regular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(-lex) regular</td>
</tr>
<tr>
<td></td>
<td>(+lex)</td>
</tr>
<tr>
<td></td>
<td>named entity</td>
</tr>
<tr>
<td></td>
<td>compound</td>
</tr>
<tr>
<td></td>
<td>idiom</td>
</tr>
<tr>
<td></td>
<td>collocation</td>
</tr>
</tbody>
</table>

Cohen's \(\kappa = 0.61\)
(significant agreement)
Evaluation experiment 2: stratified evaluation

**Corpus**
- Europarl (Koehn, 2005)

**Languages**
- French
- English
- Italian
- Spanish

**Size**
- on average, \(~3.7\) M words/language

**Methods**
- Baseline
- Syntax-based

**Evaluation**
- 50 types
- 5 levels (0-10%)
- total: 2000 types
- 2 evaluators/method

**Results**

**Annotation**
- (-gram) erroneous
- (+gram) (+lex) regular
- (-lex) named entity
- compound
- idiom
- collocation

Cohen’s \(\kappa = 0.61\)
(significant agreement)
## Annotation examples

<table>
<thead>
<tr>
<th>(-gram) erroneous</th>
<th>human development</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-lex) regular</td>
<td>next item</td>
</tr>
<tr>
<td>(+gram)</td>
<td>(+lex)</td>
</tr>
<tr>
<td>named entity</td>
<td>European Union</td>
</tr>
<tr>
<td>compound</td>
<td>point of order</td>
</tr>
<tr>
<td>idiom</td>
<td>same umbrella</td>
</tr>
<tr>
<td>collocation</td>
<td>table amendment</td>
</tr>
</tbody>
</table>
## Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>regular</th>
<th>named entity</th>
<th>collocation</th>
<th>compound</th>
<th>idiom</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>regular</strong></td>
<td>422</td>
<td>6</td>
<td>222</td>
<td>51</td>
<td>7</td>
</tr>
<tr>
<td><strong>named entity</strong></td>
<td>26</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>collocation</strong></td>
<td>315</td>
<td>63</td>
<td>11</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>compound</strong></td>
<td></td>
<td></td>
<td>64</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>idiom</strong></td>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation results

Outer ring – baseline; inner ring – syntax-based method

Statistical significance:
- +gram: $t(1436) = 26.7, p < .001$
- +lex: $t(1436) = 11, p < .001$
- collocation: $t(1436) = 9.2, p < .001$
Conclusion

- Parsing technologies, traditionally seen as inappropriate for large-scale processing of corpora, are today the main ingredient for accurate collocation extraction.

- The strong syntactic filter applied on the source text reduces the amount of data to process in the subsequent step to almost one quarter.

- Parsing is the solution to the combinatorial explosion problem in the task of identifying longer collocations in text (e.g., *be a major turning point, to stand in stark contrast*).
Thank you for your attention!
Sabine Bartsch.
2004. 
*Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Cooccurrence.*
Gunter Narr Verlag, Tübingen.

Elisabeth Breidt.
1993. 
Extraction of V-N-collocations from text corpora: A feasibility study for German.
In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 74–83, Columbus, USA.

Simon Charest, Éric Brunelle, Jean Fontaine, and Bertrand Pelletier.
2007. 
Élaboration automatique d’un dictionnaire de cooccurrences grand public.

Nick Ellis.
2008. 
Phraseology: The periphery and the heart of language.

Stefan Evert.
2004. 


Diana Lea and Moira Runcie, editors.
2002.
Oxford University Press, Oxford.

Dekang Lin.
1998.
Extracting collocations from text corpora.
In *First Workshop on Computational Terminology*, pages 57–63, Montreal, Canada.

Dekang Lin.
1999.
Automatic identification of non-compositional phrases.
In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324, Morristown, NJ, USA.

Yajuan Lü and Ming Zhou.
2004.
Collocation translation acquisition using monolingual corpora.
In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL’04)*, pages 167–174, Barcelona, Spain.

Igor Mel’čuk.
1998.
Collocations and lexical functions.

Athina Michou and Violeta Seretan.
2009.
A tool for multi-word expression extraction in Modern Greek using syntactic parsing.
Brigitte Orliac and Mike Dillinger.  
2003.  
Collocation extraction for machine translation.  

Darren Pearce.  
2002.  
A comparative evaluation of collocation extraction techniques.  

Violeta Seretan and Eric Wehrli.  
2007.  
Collocation translation based on sentence alignment and parsing.  

Violeta Seretan and Eric Wehrli.  
2009.  
Multilingual collocation extraction with a syntactic parser.  

Violeta Seretan and Eric Wehrli.  
2010.  
Extending a multilingual symbolic parser to Romanian.  

Violeta Seretan and Eric Wehrli.  
2011.  
FipsCoView: On-line visualisation of collocations extracted from multilingual parallel corpora.  
Violeta Seretan.
2011.
A collocation-driven approach to text summarization.
In Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011), pages 9–14, Montpellier, France.

John Sinclair.
Corpus, Concordance, Collocation.
Oxford University Press, Oxford.

Frank Smadja.
1993.
Retrieving collocations from text: Xtract.

María Begoña Villada Moirón.
2005.
Data-driven identification of fixed expressions and their modifiability.

Eric Wehrli, Luka Nerima, and Yves Scherrer.
2009.
Deep linguistic multilingual translation and bilingual dictionaries.

Eric Wehrli, Violeta Seretan, and Luka Nerima.
2010.
Sentence analysis and collocation identification.
Eric Wehrli.
2003.
Translation of words in context.

Eric Wehrli.
2006.
TwicPen: Hand-held scanner and translation software for non-native readers.

Eric Wehrli.
2007.
Fips, a “deep” linguistic multilingual parser.

Hau Wu and Ming Zhou.
2003.
Synonymous collocation extraction using translation information.

David Yarowsky.
1993.
One sense per collocation.