This project is partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

# D7.3: Evaluation of LT: tools and resources

*Author:* Richard Evans and Iustin Dornescu

*Affiliation:* WLV

*Date:* 31 October 2012

*Document Number:* FIRST_7.3.20121031

*Status/Version:* Approved, v1.0

*Distribution Level:* Public

| | |
|---|---|
| *Project Reference* | 287607 |
| *Project Acronym* | FIRST |
| *Project Full Title* | A Flexible Interactive Reading Support Tool |
| *Distribution Level* | Public |
| *Contractual Date of Delivery* | 31 October 2012 |
| *Actual Date of Delivery* | 31 October 2012 |
| *Document Number* | FIRST_D7.3_20121031 |
| *Status & Version* | Approved v.1.0 |
| *Number of Pages* | 40 |
| *WP Contributing to the Deliverable* | WP7 |
| *WP Task responsible* | Richard Evans |
| *Authors* | Richard Evans and Iustin Dornescu |
| *Other Contributors* | Eduard Barbu, Elena Lloret Pastor, and Paloma Moreda |
| *Reviewer* | Constantin Orăsan |

| | |
|---|---|
| *EC Project Officer* | Marco Marsella |
| *Keywords:* | Resources, evaluation, syntactic simplification, WSD, anaphora and coreference resolution |

*Abstract:*

This deliverable presents the set of tools and resources developed or acquired in order to support automatic intrinsic evaluation of the language technologies to be developed in the FIRST project. Automatic evaluation is a necessary process in software development and quality control and would be impossible in the absence of such tools and resources. The efficacy of the software when processing documents that end users seek to access has been ensured by developing the resources on the basis of corpora of three disparate genres/domains: news, patient healthcare information, and literature. The deliverable details a set of general tools and resources and describes the way in which these have been adapted in order to provide the means to assess the processing of structural complexity (WP3), the processing of ambiguity in meaning (WP4), and the generation of personalised documents/image retrieval (WP5). Statistical estimation of the consistency and reliability of the resources (via assessment of inter-annotator agreement) is also presented. Where appropriate, details of third-party tools and resources acquired for this purpose are also presented.

## Revision

| Version | Date | Revision |
|---|---|---|
| 1.0 | 31 October 2012 | Document released |

# Table of Contents

***1. Introduction*** ..... ***4***

***2. General tools and resources*** ..... ***4***

Corpora useful in the context of FIRST ..... 7

Bulgarian-language resources ..... 7

English-language domain specific corpora ..... 7

Spanish-language domain specific corpora ..... 8

***3. Structural complexity*** ..... ***8***

Bulgarian ..... 9

English ..... 9

Spanish ..... 19

***4. Ambiguity in meaning*** ..... ***20***

Coreference resolution ..... 20

English ..... 20

Bulgarian ..... 26

Spanish ..... 27

Concept tagging ..... 29

Word sense disambiguation ..... 29

Named entity disambiguation ..... 30

Image retrieval ..... 31

Figurative language ..... 32

Bulgarian ..... 32

English ..... 34

Spanish ..... 34

***Conclusions and plans for the next 12 months*** ..... ***36***

***6. References*** ..... ***37***

# 1. Introduction

The language processing components developed in WP3, WP4 and WP5 rely on various resources for training and testing. This deliverable presents the work carried out in the first year of the project in order to identify and develop the resources necessary for implementing the language technology components for English, Spanish and Bulgarian. The work produced so far tried, as much as possible, to maximise the reuse of existing resources and components, but where necessary, it created some specially for the purpose of our project.

Given the multilingual nature of the project and the fact that several difficult language processing tasks had to be addressed at the same time, it was not possible to address all the tasks in the same amount of depth in all languages. In this way, it was hoped to gain insights into how to tackle each task and acquire the know-how to satisfactorily address all the tasks by the end of the project. This is particularly relevant for Bulgarian, a language for which there is only *fragmentary support in terms of text analysis and speech and text resources*.[1]

The rest of the document presents general tools and resources which are useful for wide range of tasks, followed by sections focused on the resources used in processing structural complexity and ambiguity in meaning.

# 2. General tools and resources

This section presents general purpose tools which are employed in preprocessing text for many of the components developed in FIRST. They have the advantage that they either work out of the box or they can be easily adapted for the purpose of the project. In addition, many of them are multilingual. This section also presents textual resources available for all the languages in the project.

[1] META-NET White Paper Series: Key Results and Cross-Language Comparison http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

### *GATE*

GATE[2] is a framework for creating robust and maintainable text processing workflows used for the development of a wide variety of language technologies, including: cancer research; drug research; decision support; recruitment; web mining; information extraction; and semantic annotation. It is an integrated development environment for language processing components bundled with a comprehensive set of plugins[3] ranging from part-of-speech tagging and syntactic parsing to named entity recognising, information extraction and knowledge-based semantic information management. It also integrates and interoperates with other Natural Language Processing (NLP) systems, such as LingPipe, OpenNLP, UIMA, and others. The language processing components developed in FIRST use a document format compatible with the GATE framework, which allows easy reuse of existing tools and technologies.

### *FreeLing*

FreeLing[4] is an open source language analysis tool suite, released under the GNU General Public Licence, which implements several NLP tools, such as: morphological analysis, named entity detection, shallow -, full - and dependency parsing, word sense disambiguation and others. FreeLing is extensively used by the Spanish NLP research community, and in five years since the first version was released in 2004, it reached over 10,000 downloads, and a growing user community which has extended the three languages initially supported (Catalan, English and Spanish). Currently supported languages are Asturian, Catalan, English, Galician, Italian, Portuguese, Russian, Spanish and Welsh.

### *TreeTagger*

TreeTagger[5] is a tool for annotating text with part-of-speech and lemma information. The tagger has been successfully used to tag German, English, French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese, Swahili, Latin, Estonian and old French texts and is adaptable to other languages for which a lexicon and a manually tagged training corpus are available. TreeTagger can also be used as a chunker for English, German, and French. It provides state-of-the-art performance for Bulgarian POS tagging.

---

[2] GATE: http://gate.ac.uk/
[3] GATE plugins: http://gate.ac.uk/gate/doc/plugins.html
[4] FreeLing: http://nlp.lsi.upc.edu/freeling/
[5] TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

### *Pyramid crawler*

The Pyramid crawler was employed for building the different domain-specific corpora. This crawler was developed at the GPLSI research group[6] of the University of Alicante and it has been evaluated in different contexts (Fernández, Gómez and Martínez-Barco, 2010), showing a good performance in restricted and very specialized domains. Moreover, it is been used within real applications, such as the Virtual Observatory of Technology Transfer of the University of Alicante[7]. This crawler is not language-specific, it is used to collect corpora in Bulgarian and English as well as Spanish.

### *Weka*

Weka[8] is a collection of machine learning algorithms for data mining tasks (Witten and Frank, 2005). Moreover, Weka is a language-independent toolkit, that is very easy to run, and the input format it requires can be easily adapted from other formats. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. In addition, Weka allows analysing and testing a wide range of machine learning algorithms, as well as setting and tuning their specific parameters, with the final purpose of producing statistical models that can later be used and integrated in other applications. In FIRST the Weka toolkit is used by several NLP services such as the English syntactic processor or the Spanish pronominal coreference resolver.

### *Semantic networks*

WordNet[9] is a large lexical database of English in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. In FIRST, it is used as a thesaurus for accessing the meanings of ambiguous words in English. EuroWordNet[10] is a system of semantic networks for European languages, based on WordNet. It covers several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The individual networks are linked to an Inter-Lingual-Index, which is based on the Princeton WordNet. Via this index, the languages are interconnected so that it is possible to access similar words in the other languages via a word in one of those languages. The Spanish semantic network will be used in FIRST. BalkaNet comprises aligned semantic networks for the following Balkan

[6] http://gplsi.dlsi.ua.es/gplsi11/
[7] http://www.ovtt.org/
[8] Weka: http://www.cs.waikato.ac.nz/ml/weka/
[9] WordNet: http://wordnet.princeton.edu/
[10] EuroWordNet: http://www.illc.uva.nl/EuroWordNet/

languages: Bulgarian, Greek, Romanian, Serbian, Turkish and an extension of the Czech wordnet to EuroWordNet. The Bulgarian network, BulNet[11],will be employed in FIRST.

## Corpora useful in the context of FIRST

### Bulgarian-language resources

**Bulgarian National Reference Corpus**. This corpus consists mainly of online texts collected from the Internet and contains more than 400 000 000 tokens. 50% of the texts come from fiction, 30% from newspapers, 10% from legal texts and government texts and 10% from other genres. The Bulgarian National Reference Corpus is continuously updated with new texts.

**Frequency list**. A frequency list of the most frequent 100 000 Cyrillic words.

In addition to the resources listed above, a corpus of newswire documents in Bulgarian was compiled for FIRST at the University of Wolverhampton. The texts contained in the corpus were automatically crawled from the website of the online newspaper Vesti[12] which contains a wide variety of articles on topics such as politics, daily events and sport. At present, it contains approximately 3 000 documents and over 10 million words. Parts of this corpus have been annotated with various types of linguistic information relevant to developing NLP components in FIRST and will also be used for training unsupervised machine learning methods.

### English-language domain specific corpora

Texts of three different genres, ascertained by clinical partners to be of relevance to users with ASD, were collected for English. The genres were news, patient healthcare information, and literature. In addition to their relevance to end users, documents of these genres have also been noted to contain a broad range of obstacles to reading comprehension. News articles were obtained from the publicly accessible METER corpus (Gaisauskas et al., 2001), patient healthcare information was downloaded from *www.patient.co.uk* (96 000 documents containing more than 80 million words), and literature from the publicly accessible Gutenberg collection (*www.gutenberg.org*) was also compiled. In addition to these corpora which were employed directly in the development of the project specific resources described below, the consortium has access to a wide variety of English language resources, many of which outlined in the *Description of*

[11] BulNet: http://dcl.bas.bg/BulNet/wordnet_en.html
[12] http://vesti.bg

*Work* (see Table 4). This is due to the fact that English has better support than the rest of the languages considered in this project.[13]

**Spanish-language domain specific corpora**

A set of three domain-specific corpora in Spanish was created for the FIRST project. These corpora were collected using the Pyramid crawler described earlier and covers three domains: *newswire*, *medical* texts and *literary* documents. The newswire corpus was collected using well-known national and local newspaper sources, such as El País[14], El Mundo[15], Información[16], or La Verdad[17]. The newswire texts included in the corpus are from the year 2011, and they are not restricted to a particular section, but also provide news about general topics (politics, sports, science, etc.). This corpus is appropriate for the project, because it contains documents which are accessible to the general public. A total of 505 594 documents were collected (more than 180 million words).

The medical documents contain general information and were collected from Medline Plus[18], instead of collecting very specialized information, such as medical articles. This is justified by the nature of the project, which aims to improve reading comprehension for people with autism. In this sense, it will be more likely to read general documents about health rather than specific scientific publications. A total of 7168 documents were collected (more than 1.1 million words).

The literary corpus contains fairy tales for children[19] (121 documents), as well as electronic novels for adults[20] (335 documents). The number of documents, 456, is lower than for the newswire and medical corpora, but documents are much longer, totalling over 20 million words.

## 3. Structural complexity

A range of resources were obtained to support evaluation of language technologies developed in WP3. They are presented according to a structure in which resources supporting different LT functions are associated with each of the three languages in turn.

[13] META-NET White Paper Series: Key Results and Cross-Language Comparison (http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison)
[14] http://elpais.com/
[15] http://www.elmundo.es/
[16] http://www.diarioinformacion.com
[17] http://www.laverdad.es/
[18] http://www.nlm.nih.gov/medlineplus/spanish/
[19] http://www.grimmstories.com/es/grimm_cuentos/list
[20] http://www.gutenberg.org/browse/languages/es

### Bulgarian

**Morphological analyser.** Slovnik is a system for morphological analysis and generation. The system recognizes the word-forms of more than 110 000 Bulgarian lexemes and assigns them appropriate morpho-syntactic characteristics. A morphological analyser using a finite-state grammar approach was also developed for the CLaRK system.

**Part of speech tagger**. A neural network model for morpho-syntactic disambiguation for Bulgarian was developed under the CLaRK research programme. It achieved accuracy above 93% on a morpho-syntactic corpus consisting of 2600 sentences. TreeTagger has a part of speech tagging model for Bulgarian which achieves competitive performance.

**Syntactic Treebank.** *BulTreeBank* is a corpus containing 11 500 sentences from Bulgarian Grammar Textbooks, Newspapers, Literature and other sources of texts. The syntactic trees were created using the HPSG formalism. Several layers of linguistic information are also present: syntactic constituents, syntactic categories, head-dependency relations, coreferetial relations and ellipsis.

**Dependency TreeBank.** *BulTreeBank-DP* is a conversion of a part of BulTreeBank from the original HPSG-based annotation into a dependency format. Some information from the original encoding is omitted, such as coreferential relations and the ontological classification of named entities. Sentences containing ellipses are completely missing, due to uncertainties about the way they should be represented in the dependency format. There are 18 types of dependency relations as well as coarse-grained part-of-speech tags and morphology information.

These resources will be used to determine how syntactic information produced by statistical models can be linked to different types of syntactic obstacles. This information can be used to manipulate syntactic trees to simplify sentences which are either too long or which are structurally complex. In addition, the morphological analysers will be used to tackle lexical complexity.

### English

This section describes the work undertaken to develop the annotated resources for encoding syntactic complexity for English. This resource was developed to complement other resources of the kinds described above, but for English.

For the purpose of this project, coordination and subordination are seen as key elements of syntactic complexity. Quirk et al. (1985) define coordination as a paratactic relationship that holds between constituents at the same level of syntactic structure. The linking function occurs between conjoins that match, to a greater or lesser extent, in terms of form, function, and meaning (1). By contrast, subordination is defined as a hypotactic relationship holding between constituents at different levels of syntactic structure, referred to as superordinate and subordinate constituents (2).

(1) Amor was sentenced to nine months for each of the charges to run concurrently and told she would serve half that time in jail.
(2) Kattab, of Eccles, was not aware of the different dilutions of chloroform used to make peppermint oil.

***The annotation scheme***

In this project we assume that syntactic complexity in a text is indicated by the occurrence of particular *signs* in that text. These signs include conjunctions, complementisers, wh-words, punctuation symbols, and pairs in which a punctuation symbol is immediately followed by a conjunction, complementiser, or wh-word. These signs function either as coordinators (linking two conjoins), or subordination boundaries (delimiting the span of a subordinated constituent). The functions of coordination and bounding comprise a large number of different types. These types are denoted using 38 different class labels in the annotation scheme.

**Class labels:**

The class labels used in the annotation scheme are acronyms that abbreviate three pieces of information about the sign to which they are applied:

1. The function of the sign as either a coordinator (C), a leftmost subordination boundary (SS), or a rightmost subordination boundary (ES).
2. Information about the conjoins of coordinators and about the subordinate constituents bounded by subordination boundaries. This is:
    a. Syntactic projection level: morphemic (P); lexical (L); maximal/phrasal (M); extended/clausal (C);
    b. Syntactic category: nominal (N); verbal (V); adjectival (A); adverbial (Adv); prepositional (P); quantificational (Q)

3. Information used to distinguish between different subclasses of conjoin and subordinated constituent that involve various patterns of ellipsis (e.g. CMV2 denotes coordination of verb phrases in which the head of the right conjoin is elided whereas CMV3 denotes coordination of verb phrases in which the head of the left conjoin is elided).

There are five additional class labels used in the scheme to denote other types of coordination and subordination:

1. COMBINATORY: coordination of two conjoins to form an atomic phrase or fixed expression;
2. SPECIAL: a label used for the non-coordinating and non-subordinating functions of words that sometimes function as complementisers (e.g. specifying or anaphoric functions of *that*);
3. SSCM and ESCM: used to label the leftmost and rightmost boundaries of reported speech;
4. SSMI and ESMI: used to label the leftmost and rightmost boundaries of interjections;
5. STQ: used to label the leftmost boundaries of tag questions

To provide five examples:

- CMN1: used to label signs that coordinate noun phrases (nominal maximal projections)
- CLA: used to label signs that coordinate adjectives (adjectival lexical projections)
- CCV: used to label signs that coordinate clauses (verbal extended projections)
- SSMN: used to label signs that serve as the leftmost boundaries of subordinated noun phrases
- ESMAdvP: used to label signs that serve as the rightmost boundaries of adverbial phrases

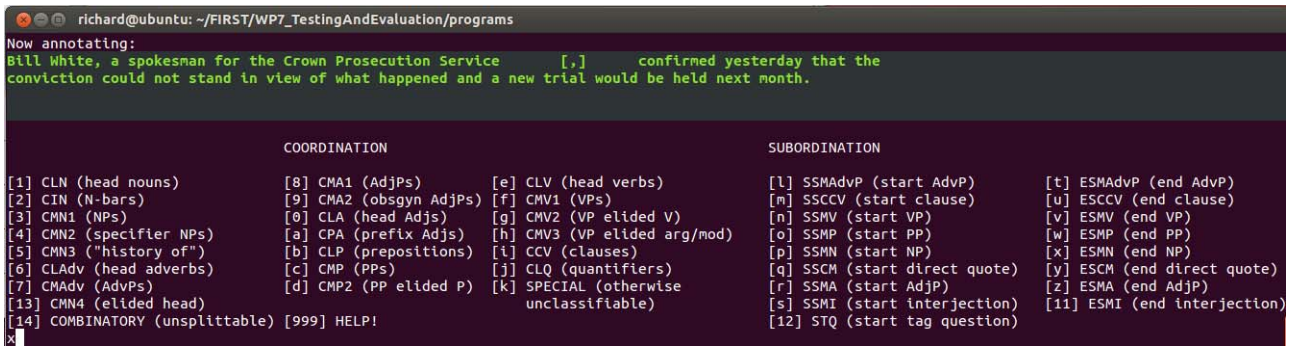


**Figure 1: The tool for manual annotation of signs of syntactic complexity**

The annotation process is facilitated by an annotation tool that has been developed (Figure 1). The tool automatically detects and highlights signs of syntactic complexity and requires annotators to select the appropriate class label for each one.

***The annotated data***

Texts from the genres of news, patient healthcare information, and literature have been annotated in accordance with the annotation scheme. Table 1 presents their characteristics.

**Table 1 Collections of documents from which resources annotated for syntactic complexity were derived**

| Source Collection | Genre | #Documents | #Sentences | #Words | #Signs of syntactic complexity | |
|---|---|---|---|---|---|---|
| | | | | | Present | Annotated |
| 1. METER Corpus | News | 674 | 22 858 | 295 718 | 30 459 | 13 312 |
| 2. www.patient.co.uk | Healthcare | 752 | 79 684 | 1 174 460 | 97 244 | 10 564 |
| 3. Gutenberg Collection | Literature | 24 | 4 468 | 113 887 | 10 551 | 10 551 |

Different numbers of the aforementioned signs of syntactic complexity were annotated in each genre with information about their class. The frequency distributions of signs and classes annotated for each of the three genres are presented in Figures 2, 3 and 4. These figures include one additional class label "OTHER" which is used to aggregate the frequencies of sign-class pairs occurring with a frequency of less than 5% of the most frequent sign-class pair.

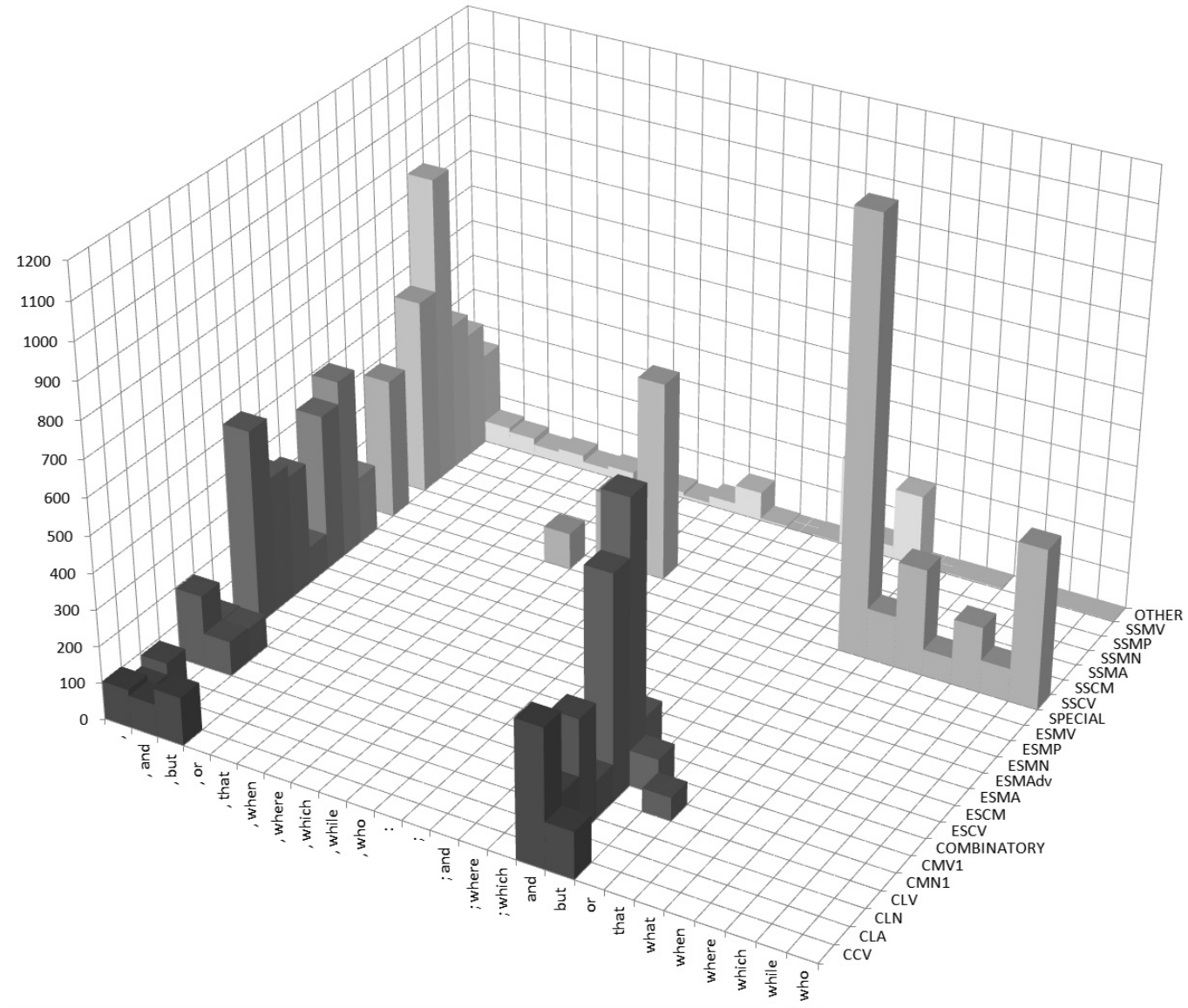


**Figure 2: Frequency distribution of signs and classes in documents from the genre of news**

In the news genre, 13312 signs of syntactic complexity were annotated. An assessment of the consistency of annotation, using Kappa, showed inter-annotator agreement at the level of 0.81, indicating “almost perfect” agreement between annotators (Viera and Garrett, 2005).

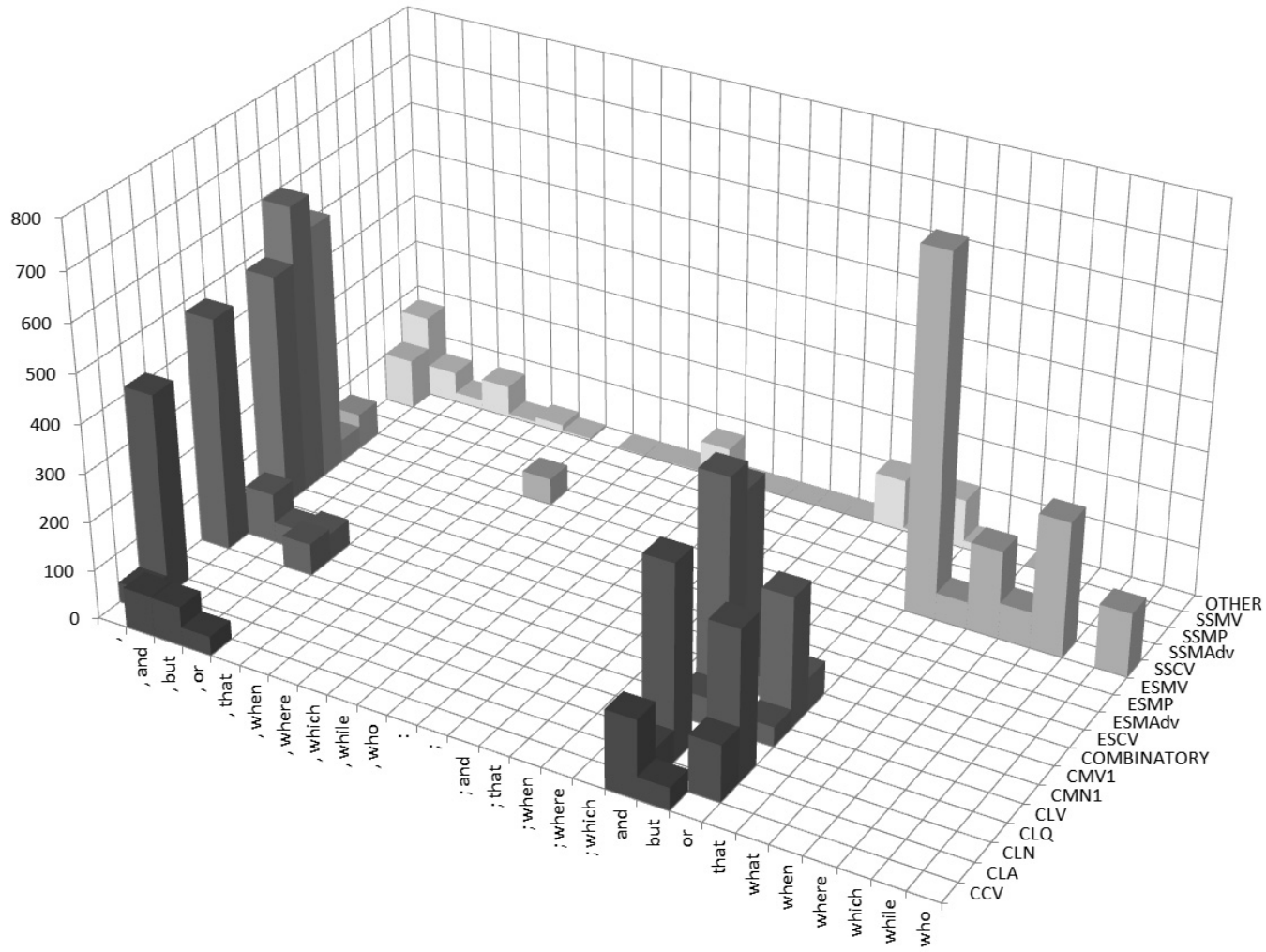


**Figure 3: Frequency distribution of signs and classes in documents from the genre of patient healthcare information**

In the genre of patient healthcare information, 10 564 signs of syntactic complexity were annotated. An assessment of the consistency of annotation, using Kappa, showed inter-annotator agreement at the level of 0.74, indicating "substantial agreement" between annotators.

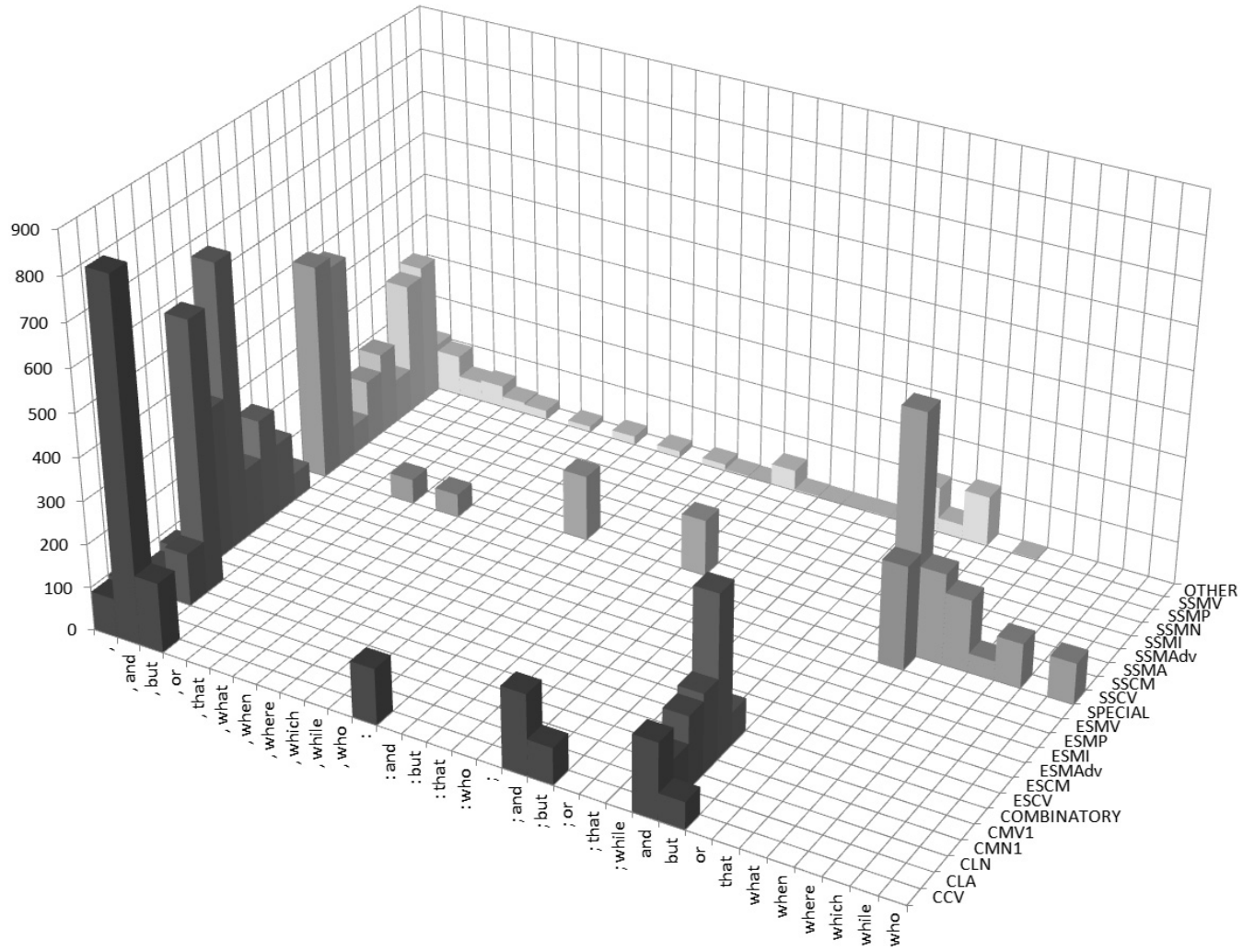


**Figure 4: Frequency distribution of signs and classes in documents from the genre of literature**

In the genre of literature, 10 551 signs of syntactic complexity were annotated. An assessment of the consistency of annotation, using Kappa, showed inter-annotator agreement at the level of 0.76, indicating "substantial agreement" between annotators.

Starting with the labelled signs of syntactic complexity and combining this with additional linguistic information such as morphological and syntactic annotations, the structure of complex sentences can be determined. Simplification operations can then be applied which will convert an input sentence containing *n* signs of syntactic complexity into two or more sentences, each of which contains at most *n-1* signs of syntactic complexity. It should be noted that while simplification of sentences containing subordinated constituents is triggered by detection of the leftmost boundaries of those constituents (the class labels

starting with SS_), some aspects of the simplification process also include detection and removal of the rightmost boundaries of those constituents.

Figure 5 displays the frequency of different classes of syntactic complexity signs occurring in documents of three genres (news, patient healthcare information, and literature). The partitions of each bar indicate the proportion of cases instantiated by different signs of syntactic complexity. Given that the development of simplification rules is labour intensive, such rules will initially be developed for the subset of most common classes and signs (Figure 6).In the project so far, considering these eleven signs (four subordination boundaries [*who, that*, *which*, *which*], three coordinators [*and*, *but*, *or*], and four that may be used with either function [**,**(comma), **;**(semicolon), ***,_and***, ***,_but***]) and ten classes to which they may belong (five coordinating classes [CCV, CMV1, CMN1, CMP, and CMA1] and five leftmost subordination boundaries [SSCV, SSMN, SSMA, SSMP, and SSMV]), 660 different rules have already been implemented. Progress in WP2 (D2.1) demonstrated that lexical coordination is less of an obstacle to reading comprehension for people with autism. For this reason, there is less emphasis on the development of simplification rules for classes CLN, CLV, CLP, CLA, CPA, CLAdv, or CLQ.

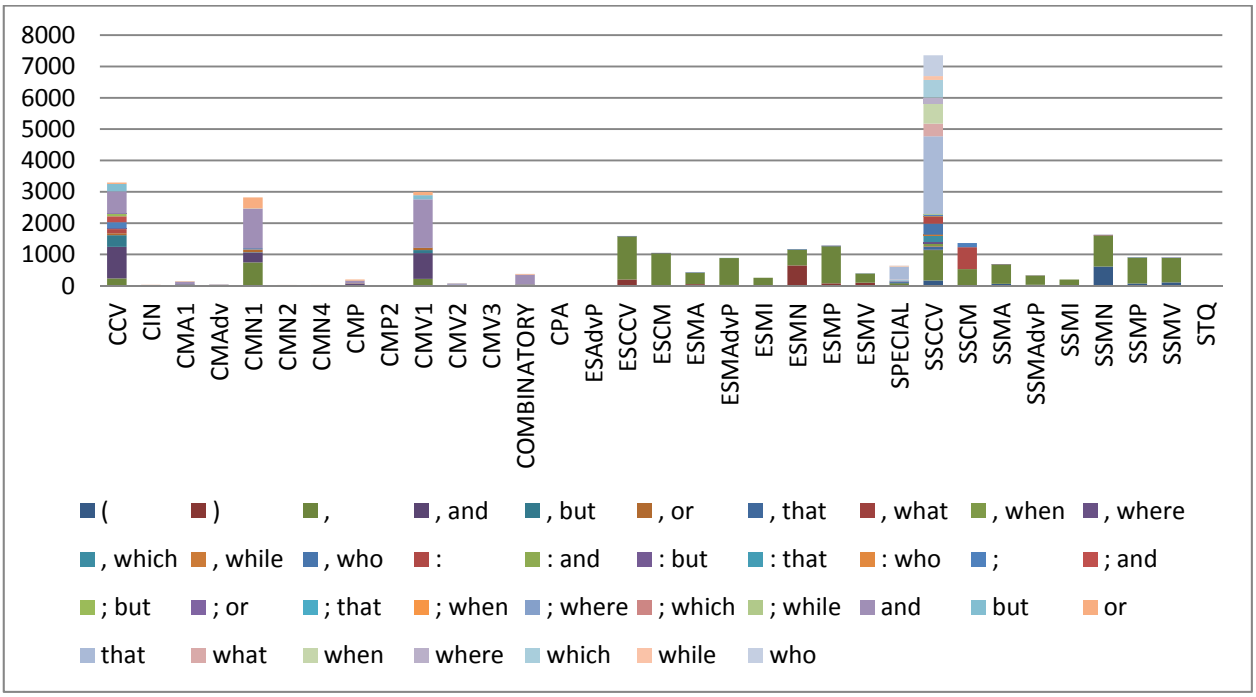


**Figure 5: Frequency distribution of signs and classes of syntactic complexity in all genres**

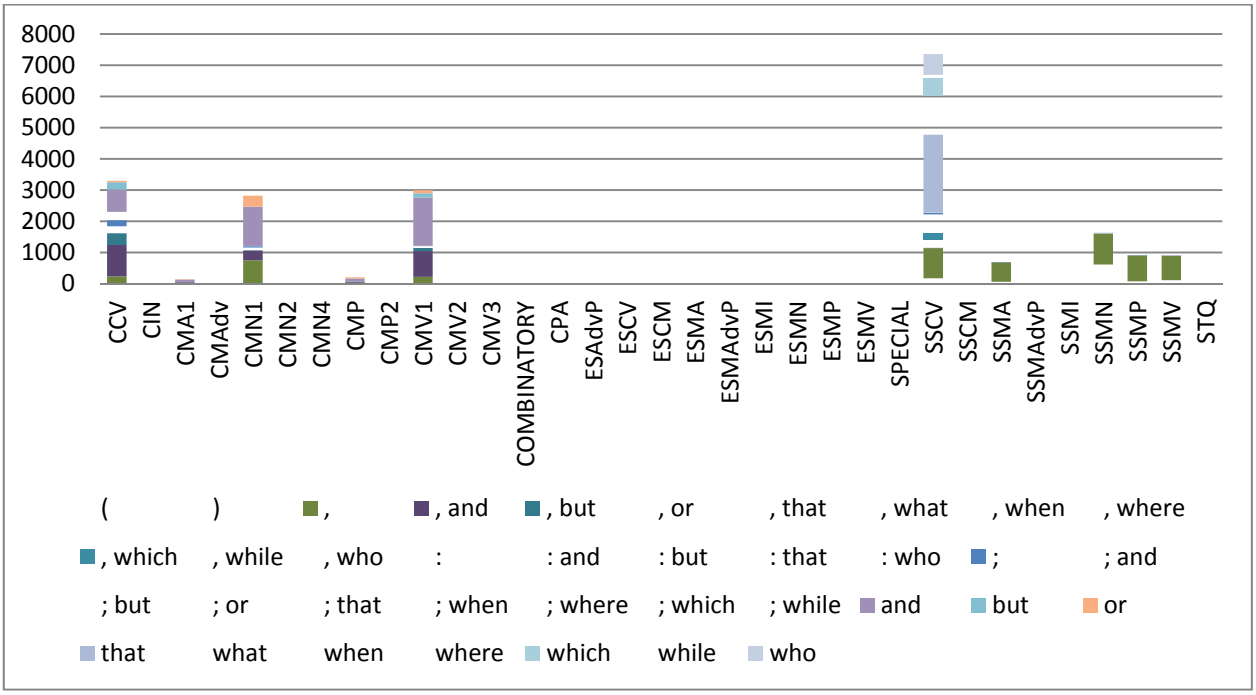


**Figure 6: Frequency distribution of signs and classes of syntactic complexity triggering automatic simplification for all genres**

**Example**

Given the sentence:

*The court was told there was no warning label on the bottle and the pharmacy's formula book was out of date and difficult to read*.

It contains two signs of syntactic complexity which are used to determine the structure of the sentence:
*The court was told* **(***there was no warning label on the bottle* **[**$^{CCV}$and**]** **(***the pharmacy's formula book was out of date***[**$^{CMA1}$and**]** *difficult to read***))**.

The difficulty in presenting this sentence is that one of the two CCV conjoins itself consists of two CMA1 conjoins. This sentence can be split into the following three simple sentences which do not contain signs of structural complexity:

1. The court was told there was no warning label on the bottle.
2. The court was told the pharmacy's formula book was out of date.
3. The court was told the pharmacy's formula book was difficult to read.

**Evaluation**. To assess the quality of the simplified sentences produced by the system, a user evaluation will be conducted in which users will indicate which of several versions of a sentence they find easier to

understand. A collection of 100 sentences processed by our software and then post-edited by human annotators was produced with the purpose of evaluating the system, but the quality and validity of this resource has not been ascertained. The editing time can be used as a proxy indicator for the quality of the simplified sentences, but such an empiric/extrinsic evaluation will rely on carers using the system, and is therefore subjective and not reproducible. Some performance metrics from the field of machine translation, such as the BLEU score, might be applicable to a certain extent.

The annotation method used to develop resources for English has been piloted for Bulgarian. It is expected that syntactic complexity (coordination and subordination) is similarly indicated by the occurrence of different signs in Bulgarian. The set of English signs has been translated and an annotation tool developed for that language. The sufficiency of these resources as a means to develop annotated resources in Bulgarian will be more fully assessed early in the second year of the project.

***Additional resources***

The automatic processing of structural complexity in English can also be enhanced via reuse of existing syntactically annotated resources such as the SUSANNE corpus (Sampson, 1995) and the Penn Treebank (Marcus et al., 1993). The Susanne Corpus is a freely available corpus developed at Oxford University. It contains a subset of the Brown Corpus (64 files belonging to 4 categories comprising a total of 130 000 words: the categories include “press reportage”; “belles letters, biography, memories”; “scientific and technical writing”; “adventure and Western fiction”). Each word in the corpus is manually annotated with information on its position, part of speech tag, orthographic form, and lemma. Of particular relevance to the project, each word is also assigned a label that encodes syntactic information about the word. Aggregating the syntactic information associated with each word enables a full syntactic analysis of the sentence to be derived. The Susanne Corpus marks 17 types of clause: main clause, subordinate clauses (adverbial, nominal, relative, comparative etc.), non-finite clauses (present participle clause, infinitival clause, for-to clause etc.) and verbless clauses (with clause, special as clause, reduced relative clause). It is expected that this information can be exploited in automatic methods for simplification of sentences including coordinated and subordinated clauses, which are the two most common types of complexity observed in documents of different genres. The Susanne corpus has already been exploited in the development of clause splitters in English (Orasan, 2000). Identifying sub-clauses is a type of shallow parsing problem which could provide additional information for simplifying complex sentences, especially when dealing with subordinated clauses. This type of approach is particularly relevant for long sentences

and, if successful, could more easily be adapted from English to Spanish and, particularly, to Bulgarian, which has fewer resources and tools.

## Spanish

**AnCora Corpus**

AnCora is a corpus of annotated Spanish and Catalan newswire documents. The annotations are at different levels, including syntactic constituents and functions. In this respect, for processing structural complexity, the identification of coordinators and subordinators is very useful, so this corpus may be appropriate to automatically learn rules concerning how such structures are created. The Spanish part of the corpus contains 500,000 words.

**UAM Spanish Treebank**

UAM Spanish Treebank is a corpus of Spanish newswire texts, where sentences have been annotated syntactically. In total, there are 1,600 sentences with this type of information. Moreover, the annotation guidelines facilitate the understanding of the annotation, thus being able to extend the size of the corpus, or to adapt the annotation guidelines for the texts analysed within the FIRST project.

**SenSem corpus**

The corpus contains more than 1 million words and consists of 25.000 sentences randomly selected from Spanish newswire articles, 100 sentences for each of the 250 most frequent Spanish verbs. It was annotated at the semantic and syntactic levels.

**FreeLing** toolkit provides two syntactic parser models for Spanish: a shallow parser and a full dependency parser called Txala. Txala parsing structure is built after sentence splitting, morphological analysis, tagging and shallow parsing have been performed. Txala builds the full syntactic trees and transforms the syntactic trees in dependency trees where each syntactic relation is labeled. The parser was evaluated on two corpora: AnCora and SenSem and scored a 73,88 % and 74,33% accuracy respectively for the task of finding the dependency head and attaching the correct syntactic relations. Unfortunately coordination and relative clauses are notoriously difficult in Spanish and the dependency parser obtains lower scores for them: between 43% and 63% accuracy, depending on the corpus.

**Connexor's Machinese** is a functional dependency partial parser which is available for Spanish and English and which was extensively used by researchers in Wolverhampton and Alicante. The dependency information provided by this tool can serve as the basis of syntactic simplification.

# 4. Ambiguity in meaning

LT resources supporting the evaluation and development of the meaning disambiguator (D4.1) are presented in this section. The meaning disambiguator includes subsystems performing coreference resolution, concept tagging, and detection of figurative language in Bulgarian, English, and Spanish texts.

## Coreference resolution

Coreference is "the act of referring to the same referent in the real world." (Mitkov, 2003: 267). Coreference resolution is the process of identifying the set of references (noun phrases, pronouns, proper names, etc.) that all refer to the same concept or entity. Current methods for coreference resolution rely on the availability of annotated resources that encode coreference between the different references that occur within samples of natural language. Such coreferentially annotated corpora enable evaluation of coreference resolution systems, which facilitates their testing and development. Research undertaken in WP2 suggests that, for readers with ASD, the most relevant aspect of coreference resolution is pronoun resolution. That is the automatic derivation of noun phrase antecedents of pronoun references. The development of resources supporting evaluation of pronoun resolution systems is therefore among the top priorities of WP7.

### English

Various resources for coreference and anaphora resolution already exist for the English language. The main research focus falls on employing these resources in order to better address the requirements of the end-users. As a starting point, the OpenBook system will use existing coreference resolution models for English. Based on user feedback it will be determined if these models need to be further adapted.

#### *Corpora*

The importance of coreference resolution for the entity/event detection task, namely identifying all mentions of entities and events in text and clustering them into equivalence classes, has been well recognized in the natural language processing community. Automatic identification of coreferential entities and events in text has been an uphill battle for several decades, partly because it can require world

knowledge which is not well-defined and partly owing to the lack of substantial annotated data. There have been several shared tasks aimed at benchmarking coreference resolution performance: MUC-6, MUC-7, ACE, SemEval-2010-Task1 and CONLL-2011.

**The ARRAU corpus**

Arrau is a corpus annotated for anaphoric relations, with information about agreement and explicit representation of multiple antecedents for ambiguous anaphoric expressions and discourse antecedents for expressions which refer to abstract entities such as events, actions and plans. The corpus contains texts from different genres: task-oriented dialogues from the Trains-91 and Trains-93 corpus, narratives from the English Pear Stories corpus, newspaper articles from the Wall Street Journal portion of the Penn Treebank, and mixed text from the Gnome corpus.

**The OntoNotes corpus**

The OntoNotes project[21] has created a large-scale, accurate corpus for general anaphoric coreference that covers entities and events not limited to noun phrases or a limited set of entity types. The coreference layer in OntoNotes constitutes one part of a multi-layer, integrated annotation of shallow semantic structure in text with high inter-annotator agreement. In addition to coreference, this data is also tagged with syntactic trees, high coverage verb and some noun propositions, partial verb and noun word senses, and 18 name entity types. The OntoNotes data distinguishes between identity coreference and appositive coreference.

**SemEval-2010-Task1**

The task is concerned with intra-document coreference resolution for six different languages: Catalan, Dutch, English, German, Italian and Spanish. The core of the task is to identify which noun phrases (NPs) in a text refer to the same discourse entity. The data provided contains coreference chains extracted from manually annotated corpora: the AnCora corpora for Catalan and Spanish, the OntoNotes and ARRAU corpora for English, the TüBa-D/Z for German, the KNACK corpus for Dutch, and the LiveMemories corpus for Italian, additionally enriched with morphological, syntactic and semantic information (such as gender, number, constituents, dependencies, predicates, etc.). Great effort has been devoted to provide the participants with a common and relatively simple data representation for all the languages.

[21] http://www.bbn.com/ontonotes/

**CONLL-2011**

The coreference resolution task used the English language portion of the OntoNotes corpus, which consists of a little over one million words from newswire (≈450k), magazine articles (≈150k), broadcast news (≈200k), broadcast conversations (≈200k) and web data (≈200k).The evaluation of coreference has been a tricky issue and there are a number of existing scoring metrics that are used to assess performance: MUC, B-CUBED, CEAF and BLANC. The evaluation used in this task was similar to that employed in SemEval-2010-Task1.

**NP4E**

NP4E[22] is a corpus encoding information about coreference relations between noun phrases and the events in which they participate in Reuters newswire articles reporting on international terrorist attacks. The corpus contains almost 55 000 words and is publicly available. Details of the annotation scheme and annotation guidelines are available from the project web page and proved to be useful in the development of annotated resources in the FIRST project.

***Tools***

A range of automatic coreference resolution systems are also available for English. They include:

**Stanford Deterministic Coreference Resolution System**

This system implements the multi-pass sieve coreference resolution (or anaphora resolution) system described in (Lee et al., 2011) and (Raghunathan et al., 2010). In the first stage, mentions are extracted and relevant information about them, e.g., gender and number, is prepared for the next step. The second stage implements the actual coreference resolution of the identified mentions. Sieves in this stage are sorted from highest to lowest precision. For example, the first sieve (i.e. highest precision) requires an exact string match between a mention and its antecedent, whereas the last one (i.e. lowest precision) implements pronominal coreference resolution. Post-processing is performed to adjust our output to the task specific constraints, e.g., removing singletons. This system was the top ranked system at the CoNLL-2011 shared task and represents the state-of-the-art performance level for English coreference resolution.

[22] http://clg.wlv.ac.uk/projects/NP4E/

**Reconcile**

Reconcile is an automatic coreference resolution system and a platform for the development of learning-based noun phrase (NP) coreference resolution systems. It achieves roughly state of the art performance on many of the most common coreference resolution test sets, such as MUC-6, MUC-7, and ACE. Reconcile utilizes supervised machine learning classifiers from the Weka toolkit, as well as other language processing tools such as the Berkeley Parser and Stanford Named Entity Recognition system.

**Other systems**

Several coreference resolution systems are currently publicly available. With regard to pronominal anaphora resolution, JavaRap (Qiu et al.,2004) is an implementation of the Lappin and Leass's (1994) Resolution of Anaphora Procedure (RAP) while MARS is an implementation of Mitkov's knowledge-poor approach (Mitkov et al., 2002), developed at WLV. GuiTAR (Poesio and Kabadjov, 2004) and BART(Versley et al., 2008) (which can be considered a successor of GuiTar) are both modular systems that target the full coreference resolution task. The toolkit was used to extend a reimplementation of the Soon et al. (2001) proposal with a variety of additional syntactic and knowledge-based features, and experiment with alternative resolution processes, preprocessing tools, and classifiers.

***Work in FIRST***

This section describes the work undertaken to annotate coreferential items within texts. In this project, the annotation encodes information on the location of referential markables such as noun phrases and pronouns and coreference between markables that refer to the same entities. This type of information can improve the accessibility of a document by enabling the user interface to indicate the correct interpretation of ambiguous phrases such as pronouns or definite descriptions to users.

As with the syntactic annotation, documents were selected from three genres: news, literature and patient healthcare information. Over all three genres, a total of 191 documents were annotated, ranging in length from around 150 words to roughly 7,000 words, with the shortest belonging to the news genre and the longest belonging to literature. Annotation was performed using the multipurpose annotation tool, PALinkA (Orasan, 2003). Figure 7 displays a screenshot of the annotation tool.

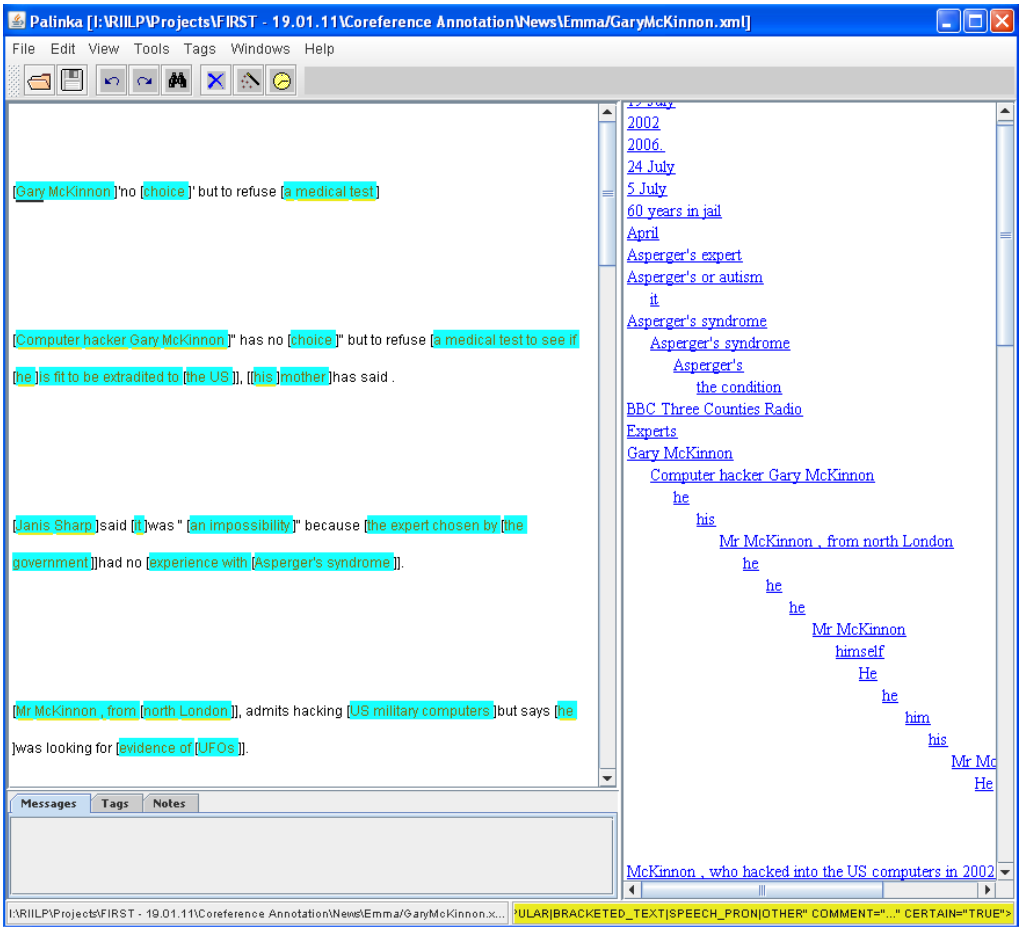


**Figure 7: A screenshot of PALinkA being used to annotate a news article with information on coreference**

The right-hand window of the interface to the annotation tool displays a link of manually selected markables. Coreferentially linked markables are indented. It will be noted that many of the NPs – often those which are indefinite –are not linked coreferentially with any others.

An example sentence is shown in (A) and a subset of the XML format used to encode coreference information is presented in (B).[23]

(A) Computer hacker Gary McKinnon "has no choice" but to refuse a medical test to see if he is fit to be extradited to the US, his mother has said.

(B) <MARKABLE ID="3"><COREF CERTAIN="TRUE" COMMENT="" ID="115" SRC="0" TYPE_REF="NP" TYPE_REL="IDENT" />**Computer hacker Gary McKinnon**</MARKABLE>**"has no**<MARKABLE ID="4">**choice**</MARKABLE>**"but to refuse**<MARKABLE ID="5">**a medical test to see**

[23] For brevity, elements such as token tags and the XML declaration are not shown in example (B).

if<MARKABLE ID="6"><COREF CERTAIN="TRUE" COMMENT="" ID="117" SRC="3" TYPE_REF="NP" TYPE_REL="IDENT"/>he</MARKABLE>is fit to be extradited to<MARKABLE ID="7">the US</MARKABLE></MARKABLE>,<MARKABLE ID="8"><MARKABLE ID="9"><COREF CERTAIN="TRUE" COMMENT="" ID="118" SRC="6" TYPE_REF="NP" TYPE_REL="IDENT" />his</MARKABLE>mother</MARKABLE> has said.

In summary, this scheme uses two tags to encode coreference between elements such as noun phrases and pronouns in text. The tags are:

1. MARKABLE: used to delimit and assign a unique ID number to each reference (noun phrase, pronoun, named entity, etc) and
2. COREF: used to indicate the unique ID number of another MARKABLE that refers to the same concept/entity as the MARKABLE to which the COREF tag is assigned.

This annotation scheme was adapted from the NP4E project. Noun phrases were marked at all levels, both definite and indefinite, from base (1) to complex (2) and coordinate (3). Embedded NPs were also identified.

(1) [Refractive errors] are [eyesight problems].
(2) [A squint] is [a condition where [the eyes] do not look together in [the same direction]].
(3) [Other treatments for [amblyopia]] include [[eye drops] and [glasses]].

Following the annotation guidelines presented by Hasler et al., (2006), a range of other elements were also markable:

- Gerunds which are true nominalisations of verbs, e.g. *[the fighting]*
- Numerals, dates and quantified NPs, e.g. *[twenty men]*
- Possessive pronouns and other possessors, e.g. *[[Kabila's] forces]*
- Interrogative pronouns functioning as possessives, e.g. *[President Alberto Fujimori, [[whose] brother Pedro] is…*
- Reciprocal pronouns, e.g. *[each other]*.

Table 2 English coreference corpora annotated for FIRST summarises the characteristics of the English coreferentially annotated corpora developed in the project.

Table 2 English coreference corpora annotated for FIRST

| *GENRE* | *UNITS* | *NUMBER* |
|---|---|---|
| News | Documents | 74 |
| | Elements (MARKABLE) | 8711 |
| | Relations (COREF) | 2670 |
| | Words | 35697 |
| Patient healthcare | Documents | 76 |
| | Elements (MARKABLE) | 32264 |
| | Relations (COREF) | 7369 |
| | Words | 169129 |
| Literature | Documents | 41 |
| | Elements (MARKABLE) | 22181 |
| | Relations (COREF) | 12254 |
| | Words | 113887 |

### Bulgarian

The methodology used to develop coreferentially annotated resources in Bulgarian is adapted from the approach applied in development of the resources for English. As in that case, the annotation scheme presented in Hasler et al (2006) was applied to Bulgarian texts using the multipurpose annotation tool, PALinkA (Orasan, 2003).

The annotation process for Bulgarian differs from that for English due to the frequent occurrence of zero-pronouns in the former language. This fact motivated the addition of an additional tag to mark the position of zero pronouns in Bulgarian texts. As in the case for explicit MARKABLE elements, coreferential relations between gaps and other markables is asserted using COREF tags (see the section on coreferentially annotated resources for English).

So far, documents belonging to three different genres have been coreferentially annotated in Bulgarian. Unlike the situation in English, these texts are aimed at young readers and are from a variety of sources:

21 newspaper articles, 16 examples of literature, and 53 documents providing healthcare information (including general health (18), mental health (17), and sexual health (18)). In total, these annotated resources comprise 68 000 words, encoding 3 337 coreference relations between 15 300 different markables.

## Spanish

### SemEval 2010 corpus

The SemEval-2010 corpus for Task *1: Coreference Resolution in Multiple Languages* (Recasens et al., 2010) was employed to support evaluation of the pronominal anaphora resolution module, since it consists of textual data annotated with coreference information (including pronominal anaphora) of the newswire domain. Therefore, this corpus fits very well within the scope and objectives of the project. The corpus is available at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2011T01 , and Table 3 shows its statistics.

**Table 3 Properties of SemEval-2010 Task 1 Spanish corpus**

| | |
|---|---|
| Total number of documents in the corpus | **1183** |
| Number of documents included in the training set | **857** |
| Number of documents included in the development set | **140** |
| Number of documents in included the test set | **168** |
| Average length of documents (number of words) | **270** |
| Number of pronouns found in the corpus | **8886** |

### Pronominal Anaphora Resolution

This module was developed within WP4, and its main aim is to detect pronouns in documents and resolve them correctly, by identifying the antecedents that they refer to. The module was tested against the SemEval-2010 corpus. The distribution of pronouns found in this corpus is presented in Figure 8.

In this deliverable, intrinsic evaluation is reported, since we focus on the evaluation of the pronominal anaphora resolution tool itself, and not on how it can help to people with autism when they use it (extrinsic evaluation will be performed at a later stage in the project). The intrinsic type of evaluation was conducted over different kinds of pronouns (personals, demonstratives, indefinites, interrogatives, relatives, and clitics). A particularity for Spanish is that clitic pronouns can appear either before or after the verb (e.g., ***te****obsesionas* or *obsesionar****te***), so we differentiate this fact in the evaluation.

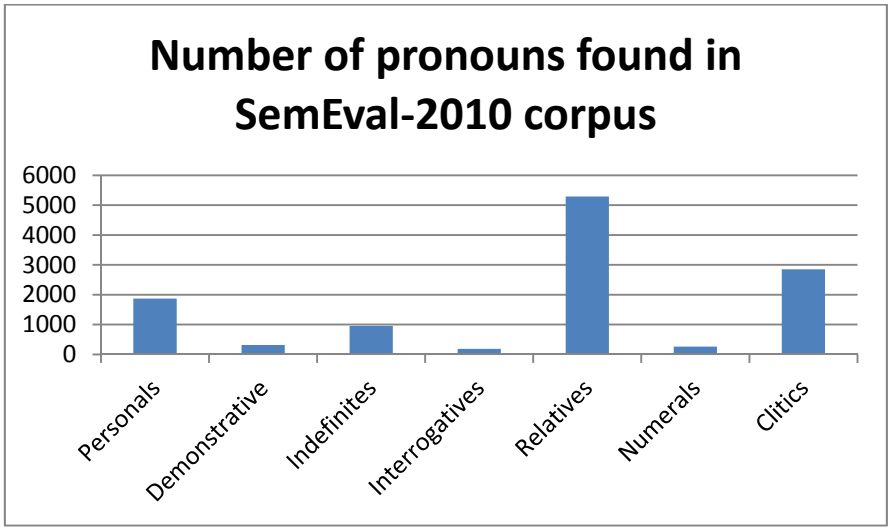


**Figure 8: Distribution of pronouns found in SemEval-2010 Spanish corpus**

For the evaluation, the test part of SemEval-2010 corpus was employed and the standard evaluation metrics of precision, recall and F-measure commonly used in Natural Language Processing research area were computed. However, we only focused on the precision value, since what it is important in this task is how many pronouns are correctly resolved.

Several statistical models were built using Weka and the best-performing machine learning algorithm was *Voted Feature Interval* (VFI) which achieved an average precision of 69.70%. This result compares favorably with state-of-the-art performance (state-of-the-art results in coreference resolution for Spanish are around 58% (Recasens, 2010)). More specifically, for pronominal anaphora resolution, results range between 45% and 65% for rule-based systems (Ferrández and Peral, 2003; Mitkov et al., 2007). To the best of our knowledge, there is no previous work applying machine learning for pronominal anaphora resolution in Spanish; however, for other languages, such as Basque or Norwegian, results for this type of coreference and using machine learning algorithms are 65.30% (Agerri et al., 2010) and 67.50% (Noklestad, 2009), respectively.

An example of a text fragment[24] which contains two personal pronouns (**ella**) that were correctly resolved by our pronominal anaphora resolution module (its antecedent is underlined), is presented below.

[24]The whole article can be found at http://www.elpais.com/pda/index.php?module=elp_pdapsp&page=elp_pda_noticia&idNoticia=20121003elpneppol_4.Tes&seccion=pol

*[…]Para*Soraya Rodríguez*, todo eso muestra una actitud "dictatorial" y de falta de respeto al Parlamento.* ***Ella*** *es una abogada y política española, y actualmente* ***ella*** *es portavoz del PSOE en el Congreso de los Diputados.[…]*

## Concept tagging

### Word sense disambiguation

Word Sense Disambiguation (WSD) is a key technology that automatically chooses the intended sense of a word in context. Word Sense Disambiguation will be used in both WP4 and WP5 to automatically disambiguate relevant concepts. The software written in the project will adapt already available tools that perform WSD to the project needs.

The challenge of WSD can be met via a supervised or an unsupervised approach. Supervised WSD systems have been found to be best performing in public evaluations (Palmer et al., 2001; Snyder and Palmer, 2004) but they need large amounts of hand-tagged data, which is typically very expensive to build. Given the relatively small amount of training data available, current state-of-the-art systems offer only marginal improvements over the simple "most frequent sense" (MFS) baseline.[25] As an alternative to supervised systems, knowledge-based WSD systems exploit the information present in a lexical knowledge base (LKB) to perform WSD, without using any additional corpus evidence. The main advantage of the unsupervised approach is that it is domain independent, making it suitable for this project.

The task of WSD in Bulgarian can exploited BulSemCor[26], which is a 75 000-word sense-annotated corpus of Bulgarian consisting of 500 excerpts of text, each of which contains at least 100 words. Each lexical item (simple or compound word) was manually assigned its most relevant semantic meaning from the Bulgarian wordnet, BulNet. The sysnets are linked to BalkaNet and WordNet. FreeLing[27]provides two WSD methods for English and Spanish. The knowledge-based WSD evaluated for the project is based in the use of WordNet (Miller et al., 1991) and PageRank algorithm (Brin and Page, 1998), and it was proposed by Agirre and Soroa (2009). The authors measured the performance of the algorithm using the

[25] This baseline consists of tagging all occurrences in the test data with the sense of the Word that occurs more often in the training data.
[26] Description: http://dcl.bas.bg/en/corpora_en.html, ELRA Catalog Reference number: ELRA-U-W 0129
[27] http://nlp.lsi.upc.edu/freeling/

English corpora Senseval-2 and Senseval-3 and with the Spanish corpus Semeval07. The method has two main options “ppr” and “ppr_w2w”. The first disambiguates target words individually, while the second uses context words to disambiguate the target words. The performance of the algorithm is presented in Table 4:

Table 4 WSD results on Senseval-2, Senseval-3 and Spanish Semeval07

| **Senseval-2** | | |
|---|---|---|
| **LKB** | **Method** | **Recall** |
| WordNet3.0+gloss | Ppr | 53.5 |
| WordNet3.0+gloss | ppr_w2w | 55.8 |
| - | MFS | 60.1 |
| **Senseval-3** | | |
| **LKB** | **Method** | **Recall** |
| WordNet3.0+gloss | Ppr | 48.5 |
| WordNet3.0+gloss | ppr_w2w | 51.6 |
| - | MFS | 62.3 |
| **Spanish Semeval07** | | |
| **LKB** | **Method** | **Accuracy** |
| Spanish Wnet + Xnet | Ppr | 78.4 |
| Spanish Wnet + Xnet | ppr_w2w | 79.3 |
| - | MFS | 84.6 |

Unfortunately, the unsupervised algorithm is not superior to MFS when applied in the open domain but it may be feasible to adapt it to the particular domains (e.g. news and medical texts) used in the FIRST project.

### Named entity disambiguation

Named-entity recognition (NER), also known as entity identification and entity extraction, is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, temporal expressions, quantities, monetary values, percentages, etc. NER is an important process in enabling semantic networks to be extended with

information about proper names and acronyms. This can contribute to tasks such as coreference resolution by enabling semantic agreement constraints to be enforced.

The FreeLing suite implements two methods for performing NER. The first one called “basic” uses linguistic rules to identify entities. This method is easily adaptable to other languages and its estimated performance is about 85% (correctly recognized named entities).
The second method is based on machine learning and is called “bio”. It has a higher precision, over 90%, but it is slower than the “basic” method, and adaptation to new languages requires a training corpus plus some feature engineering. Both methods are already implemented for English and Spanish and adaptation of the first method to Bulgarian is feasible.

In the FIRST project we will use NER as a step towards Named Entity Disambiguation. After the entities are disambiguated they can be linked with definitions in Wikipedia and dictionaries or illustrated by images thus facilitating their comprehension by the system users.
GATE contains information extraction components which can identify named entities in English, Spanish and Bulgarian. It also provides a variety of gazetteer types as well as ontology-based semantic information management plugins which enable entity mentions in text documents to be linked to entries in large knowledge bases.

**Image retrieval**

Difficult concepts can be illustrated using corresponding images to facilitate reading comprehension by users of our system. The image corpus used in FIRST, ImageNet, is described in more detail in in deliverable D5.i. To identify suitable images, the method exploits information arising from both WSD and NER processing (e.g. annotations produced by modules implemented in FreeLing and described in a preceding section).

A preliminary qualitative evaluation was made of the corpus of documents used in SemEval 2010. The main motivation for this evaluation was to assess the efficacy of combining information from WSD and ImageNet. In particular we wanted to understand how well the WSD algorithm performs for the concepts to be illustrated by images and whether or not the retrieved images are pertinent to the document context. To perform the evaluation, we annotated (with the offline software supporting the Web Service) 354 documents representing news articles in English. The number of token images retrieved for the corpus was 9134 with an average number of 26 images per document. We randomly chose 3 documents and evaluated

the accuracy of WSD for the concepts illustrated by images. We also assessed the relevance of the images for the overall context of the document.

The precision of WSD was 76%, and it was computed as the number of correctly disambiguated words (concepts to be illustrated by images) divided by the total number of concepts to be illustrated by images. As expected the calculated WSD precision is in line with the numbers reported by FreeLing authors.
For the correctly disambiguated words, 40% of the retrieved images were deemed relevant to the context of the document. The main reason for this is best understood by the following sentence:

*He is the President of Serbia.*

The concept *president* is disambiguated to the correct WordNet sense: (the chief executive of a republic) but the actual image retrieved is not that of the president of Serbia. It correctly illustrates the idea of a president in general but fails to represent the contextual term: *President of Serbia*.
In subsequent months, we will try to improve the relevance to context by making a better linguistic analysis. This analysis can be performed using a parser and/or identifying the relevant terms and entities in a document before performing WSD. The disambiguation should then be performed only for whole terms or entity names and not for their components.

## Figurative language

Detection of figurative language in Bulgarian, English, and Spanish will be based on a common approach relying on dictionary look-up and flexible matching algorithms capable of recognising the occurrence of both canonical and variant forms of those expressions in different texts. The acquisition of dictionary resources is language specific.

### Bulgarian

Resources related to figurative language accumulated so far for Bulgarian consists of three main areas:

#### *Metaphors*

People with autism experience certain difficulties with figurative language but like all other readers they are active learners as well. In order to support not only their reading comprehension but also their language development, we compiled a list of metaphors, which can be expressed using a wider range of phrases than idioms (and can easily be extracted from already existing phraseological dictionaries) but are often seen as set phrases, known as “conventionalized” metaphors (Lakoff and Johnson, 2003).

At this stage the list consists of 310 metaphorical expressions derived from newspaper articles, Wikipedia articles and Internet blogs. Examples which have their similar analogues in English are: "главанасемейство" ("head of the family"); *давамдума* ("to give my word"); *езиковабариера*("language barrier"); *жълтатапреса* ("yellow press"); *крокодилскисълзи* ("crocodile tears"); *майчинезик* ("mother tongue"); *остареакция* ("sharp reaction");*полагамосновитена* ("to lay the foundations of"),etc. as well as expressions which signify particular persons, events, geographical locations or notions: *Страната на Изгряващото слънце*– The Land of The Rising Sun; *Желязната лейди*– The Iron Lady; *Желязната завеса*– The Iron Curtain;*Новият свят* – The New World, etc.

***Expressions with figurative meaning signifying feelings and emotions***

As emotion recognition deficit is one of the core problems in ASD, and it can be assumed that the comprehension of figurative expressions related to emotions and states of mind would be more difficult for these readers. This is due to the fact that these utterances are both figurative in meaning and refer to the problematic area of feelings. Lists of figurative expressions of emotion such as "*to feel blue*" are envisaged to be useful in identifying phrases in the document that may pose obstacles to reading comprehension. Information on the type of emotion expressed is likely to be useful in facilitating comprehension of such phrases. A gazetteer was compiled that includes entries belonging to four main emotional categories: Joy, Anger, Pain and Fear. Examples include: *извън кожата си*("out of one's skin"); *летящ* („flying"); *обезумял* („out of mind"), *трептящ*("fluttering"); умиротворен ("in peace with himself"), etc.

***Prepositions***

Another of the main areas of difficulty for people with ASD involves the comprehension and use of prepositions (Oliver, S., 1998, „Understanding Autism". Oxford Brookes University, 1998).
In combination with other words many of them often form collocations with figurative meaning. Lists containing examples of the figurative use of the most common prepositions in Bulgarian were automatically extracted from texts in the genres of literature and news. They include: *Подносана* ("under one's nose"); *Подсянкатана* ("in one's shadow");*В гнева* ("in anger"); *Под съмнение*("under suspicion"), *Междунас* ("between you and me"),*В иметона* ("in the name of"),*В отчаяние* ("in despair"), etc. The list of prepositions forming figurative collocations is derived from classic Bulgarian novels and short stories.

## English

Several resources have been identified to facilitate development of dictionary resources for use in the automatic detection of figurative language in English. These include:

- the Conceptual Metaphor database (http://cogsci.berkeley.edu/lakoff/MetaphorHome.html ), an online collection of 208 metaphors indexed by metaphor name, source domain (metaphorical meaning), and target domain (literal meaning).
- An Asperger Dictionary of Everyday Expressions (Stuart-Hamilton, 2004): a dictionary for use by people with Asperger Syndrome that defines 5 000 commonly used everyday expressions with non-literal meaning.
- The Oxford Dictionary of Idioms: a dictionary defining 5 000 English idioms.

## Spanish

### *Dictionary of metaphors*

The dictionary of Spanish metaphors developed in the first year of the project contains 100 common Spanish idioms gathered from the Internet. Each idiom is paired with an explanation of its meaning. For instance, the expression “tirar la toalla” (English: to throw in the towel) means “to give up”, and this is an example of idiom widely used in the newswire domain, specifically in political or sport news. A fragment from such a dictionary can be seen below.

*Poner en tela de juicio.* | *Expresa dudas sobre la legalidad o certeza, incluso éxito de una cosa.*
*Tirar la toalla.* | *Resignarse, dejar de hacer algo. Llevaba estudiando 5 meses para el examen, y volvió a suspenderlo. Se quedó tan decepcionada, que al final tiró la toalla y no se presentó una 6ª vez.*
*Tomar el pelo.* | *Intentar bromear o engañar a alguien. No me creo que a tu amigo le hayan salido 3 brazos, me estás tomando el pelo.*
*Tomar por el pito del sereno.* | *No tomar a alguien en serio. Tus hijos no te hacen caso. Cuando les dices que vengan a comer inmediatamente, nunca obedecen y siempre tardan un buen rato en aparecer. Te toman por el pito del sereno.*
*Ponerse las pilas.* | *Empezar a actuar ante una situación problemática y recuperar el tiempo perdido. Estamos en el tercer trimestre, y los exámenes no los llevas muy bien. Ya te puedes poner las pilas, o no aprobarás a final de curso.*
*Por si las moscas.* | *Por si acaso. No me gustan esas nubes. Voy a coger el paraguas, por si las moscas.*
*Por un tubo.* | *En abundancia, en grandes cantidades. La 4ª pantalla del juego era realmente complicada, no paraban de aparecer enemigos por un tubo.*

*Romper una lanza a favor de. | Confiar en algo o alguien. El equipo no está jugando tan mal. Rompo una lanza a favor de ellos, estoy seguro de que acabarán ganando el partido.*

### *Metaphor Identification Module*

The dictionary of metaphors is under construction, and the evaluation performed so far consisted in analyzing the frequency of appearance of such idioms in the newswire corpus also created for the project that was previously explained. In this corpus, composed of 505 594 news documents, 1922 idioms were found. Table 5 shows the 4-most frequent idioms and the number of news articles in which they appear.

**Table 5 Frequent idioms in Spanish**

| Idiom | Number of news articles it appears in |
|---|---|
| Tirar la toalla<br>(English translation: *throw in the towel*) | 352 |
| Meter la pata<br>(English translation: *to put one's foot in it*) | 204 |
| Poner la mano en el fuego<br>(English translation: *to risk somebody's neck*) | 194 |
| Marear la perdiz<br>(English translation: *to mess about*) | 189 |

The example below, taken from a news reported on the 16th of October 2011[28], contains one of these frequently occurring Spanish idioms, i.e., "*tirar la toalla*".

*[…]Pero la primera secretaria en excedencia del Partido Socialista no se resigna a* ***tirar la toalla*** *y aún confía en que su programa ortodoxo y su imagen de dirigente tenaz le deparen, a última hora, los apoyos de muchos de los seguidores de Montebourg: ese ídolo de la antiglobalización que ocupó sorpresivamente el tercer puesto el pasado 8-O, con un 17% del escrutinio.[…]*

[28] The whole text can be found at http://www.elmundo.es/elmundo/2011/10/16/internacional/1318760694.html

## Conclusions and plans for the next 12 months

This deliverable has presented the main tools and resources acquired and developed in the first year of the project for Bulgarian, English and Spanish. Over the coming 12 months the work related to testing and evaluation of language technology will continue the development and enhancement of these resources. However, it is considered that most of the necessary resources are already in place for the tasks to be addressed in the project and therefore the focus of the work will shift to evaluation and error analysis of the language technologies.

## 6. References

Agirre, E. and Soroa, A. (2009) Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (EACL '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 33-41.

Brin, S., and Page, L. (1998) The anatomy of large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7).

Fernández, J., Gómez, J.M. and Martínez-Barco, P. (2010). Evaluación de sistemas de recuperación de información web sobre dominios restringidos. *Procesamiento de Lenguaje Natural*, 45(0):273–276.

Gaizauskas, R., Foster and, J., Wilks, Y., Arundel, J., Clough, P. and Piao, S. (2001) The Meter Corpus: A Corpus for Analysing Journalistic Text Reuse, In *Proceedings of Corpus Linguistics 2001 Conference*, pp. 214-223.

Grover, C., Matheson, C., Mikheev, A., and Moens, M. (2000) LT TTT—a flexible tokenisation tool. In *LREC 2000—Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 1147–1154.

Hasler, L., Orasan, C. and Naumann, K. (2006) NPs for Events: Experiments in coreference annotation. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC2006)*, Genoa, Italy, 24-26 May, 1167-1172

Lakoff, G.and Johnson, M. (2003) *Metaphors we live by.* The University of Chicago Press, London.

Lappin, S. and Leass, H. J. (1994). *An Algorithm for Pronominal Anaphora Resolution*. Computational Linguistics, 20(4), 535-561.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M. and Jurafsky, D. (2011) Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Proceedings of the CoNLL-2011 Shared Task*.

Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B. (1993) *Building a large annotated corpus of English: the Penn Treebank*, Computational. Linguistics, 19(2), 313-330, MIT Press, Cambridge, USA.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., and Tengi, R. (1991) Five papers on wordnet. *Special Issue of the International Jornal of Lexicography*, 3(4):235-312

Mitkov, R., Evans, R. and Orasan, C. (2002): A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of CICLing-2000*, Mexico City, Mexico.

Mitkov, R. (Ed). (2003). *Oxford Handbook of Computational Linguistics*. Oxford University Press.

Mitkov, R. and Hallett, C. (2007) *Comparing pronoun resolution algorithms*. Computational Intelligence, 23 (2), 262-297

Oliver, S. (1998) *Understanding Autism*. Oxford Brookes University, 1998

Orasan, C. (2000) A hybrid method for clause splitting in unrestricted English texts. In *Proceedings of ACIDCA '2000*, Corpora and Natural Language Processing, March 22-24, Monastir, Tunisia, pp. 129 – 134

Orasan, C. (2003) PALinkA: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop on Discourse and Dialog*, Sapporo, Japan, 5 – 6 July, pp. 39 – 43

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A comprehensive grammar of the English language*, Longman.

Padró, L., Reese, S., Aguirre, E., Soroa, A. (2010) Global WordNet Conference. In *Proceedings of the 5th Global WordNet Conference,* pp. 99-105, Mumbai.

Palmer, M., Fellbaum, C., Cotton, S., Delfs, L., and Dang, H. T. (2001) English tasks: All-words and verb lexical sample. In *Proc. Of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Tolouse, France.

Peral, J., and Ferrandez, A. (2003) *Translation of Pronominal Anaphora between English and Spanish: Discrepancies and Evaluation*. Journal of Artificial Intelligence Research, Vol. 18, pages 117-147, Morgan Kaufmann.

Poesio, M. and Kabadjov, M. (2004) A General-Purpose, off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

Poesio, M. and Artstein, R. (2008) Anaphoric Annotation in the ARRAU Corpus, In *Proceedings of LREC'08*, Morrocco.

Qiu, L., Kan, M. Y. and Chua, T. S. (2004) A Public Reference Implementation of the RAP Anaphora Resolution Algorithm. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Vol. I, pp. 291-294.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D. and Manning, C. (2010) A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of EMNLP-2010*, Boston, USA.

Recasens, M., Màrquez, Ll., Sapena, E., Martí, M.A., Taulé, M., Hoste, V., Poesio M. and Versley, Y. (2010). SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, Association for Computational Linguistics, USA, 1-8.

Recasens, M. and Hovy, E. (2010) Coreference Resolution across Corpora: Languages, Coding Schemes, and Preprocessing Information. In *Proceedings of ACL 2010*, Uppsala, Sweden.

Sampson, G. (1995) *English for the computer: the SUSANNE corpus and analytic scheme*. Oxford University Press.

Snyder, B., and Palmer, M. (2004) The English all-words task. In *ACL 2004 Senseval-3 Workshop*, Barcelona, Spain.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001) *A machine learning approach to coreference resolution of noun phrases. Computational Linguistics*, 27(4):521–544.

Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E., Buttler, D. and Hysom, D. (2010) Reconcile: A Coreference Resolution Platform, Tech Report - Cornell University.

Stuart-Hamilton, I. (2004) *An Asperger Dictionary of Everyday Expressions*. Jessica Kingsley Publishers, London.

Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A. (2008) BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*

Viera, A.J. and Garrett, J.M. (2005) Understanding inter-observer agreement: the kappa statistic. *Fam Med.*, 37(5):360-3.

Witten, I. H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.