

Ambiguity and Contradiction From a Morpho-Syntactic Prototype Perspective

M. D. López De Luise - mlopez74@palermo.edu
Department of Informatics Engineering , Universidad de Palermo University
Av. Córdoba 3501, Capital Federal, C1188AAB, Argentina

Abstract— In this paper, the contradiction and ambiguity problems in Natural language Processing are briefly introduced. We also present the morpho-syntactic WIH (Web Intelligent Handler) prototype and the overall approach it takes to process any Spanish text. Finally, we analyze how it processes Spanish sentences with contradictions or ambiguities using its own perspective, despite deeper linguistic considerations.

Index Term — Text-Mining, Automatic summarization, morphosyntactic analysis.

I INTRODUCTION

Communication through written texts has a certain coherence [1] and structure[2] according to the type of text. When text is coherent, there is a unique sense in it (named intra-textual coherence). But there is a conceptual coherence that exceeds the text. It is inherent of the culture and historical situation. This paper does not study any kind of coherence. It just analyzes parts of texts. In the following sections, the contradiction and ambiguity linguistic concepts are introduced along with some of the state of the art in that area. Afterwards, WIH (Web Intelligent Handler) prototype is presented and tested in several test cases with contradiction and ambiguity.

II CONTRADICTION

Contradiction has been studied since Aristotle [3]. In general, it is considered as a concept:

-Logical or analytic: as a reasoning distortion. One example is the Aristotle no contradiction principle.

-Dialectic or synthetic: in a philosophic context, as an expression with its own essence. Unamuno Kierkegaard and Hegel had done part of this work.

From the computational perspective, only the first alternative can be considered. Work done with contradiction could be for Natural Language Processing frameworks [4], Distributed processing, Knowledge Based Systems [5] [6] [7] [8], etc. Some working prototypes are ASKER [9], CLASSIC [10], ECD [11], and XLE [4].

III AMBIGUITY

Written text is mostly disambiguated by the reader knowledge of its context and its main topic [12] [13]. But ambiguity is intrinsic to natural Languages morphology, syntax and semantics [2]. They can be classified as:

Lexical ambiguity: it is generated by the polysemy of a word.

Syntactic ambiguity: due the combination of the syntactic structure with certain nesting and embedding clauses.

Semantic ambiguity: originated by the logical structure of a sentence (not related to subordinated sentences as in the previous case).

Anaphora: originated by textual components without semantics. But they are syntactically right.

Some of the work in this area is related to lexical [14] [15], syntactic [16], morphologic [17], phonemic [18] [19], semantic [20] [21] levels or certain combination of them [22] [23]. In other prototypes as CLICK-TALP this task is performed manually [24] [25].

IV WIH PROTOTYPE

The Web Intelligent Handler (WIH) is a partially working prototype to process Spanish based web content in order to generate a meta-web for browsing and querying support. It was first introduced in [26]. It has a three-layered design (see Fig. 1):

-Internal Structure: gets data and metadata from the WWW and processes it to derive a set of Homogenized Basic Elements (HBE). These elements constitute a representation in an internal language.

-Virtual Structure: processes the actual stream of HBE and makes a structure named E_{ci} for the name in Spanish Estructura de Composición Interna, Internal Composition Structure in English. An E_{ci} is an oriented graph representing a statement in the original text. Sets of E_{ci} are then processed to make an E_{ce} (for the name in Spanish Estructura de Composición Externa, External Composition Structure in English, a supra-structure composed by a set of E_{ci} structures, all of them related to the same text).

-Visual Structure: it works with a Virtual Network composed by the set of E_{ci} and E_{ce} structures. It can be considered as an interface between the Virtual Structure and any user.

The WIH prototype has implemented a first version of the Internal Structure and Virtual Structure. In the Virtual Structure, there is a set of components to perform: the composition of E_{ci} and E_{ce} structures (Composition Engine, CE), insertion into the Virtual Net (Assimilation Engine, AE) through a set of functions (named effect functions, f_e) regulated by a set of dynamic parameters (metric functions, f_m). All the activity is controlled by a feedback system (composed by a general controller System Controller, a Manager for f_e named Metrics Manager, and a manager for f_m named Metrics Engine).

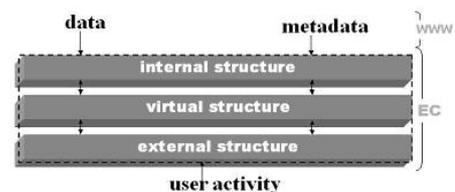


Fig. 1. Three layered structure of the EC. Data and metadata are extracted from the WWW.

The system performs all the activity through a set of f_c that can be changed dynamically. Some of such effect functions define the way a set of HBE is transformed to an E_{ci} , and how a set of E_{ci} is converted into an E_{ce} . For this activity, a metric p_o was defined as a set table to categorize each HBE and therefore determine how to process it. In general terms E_{ci} weight is in $p_o^{E_{ci}} \in [-1, +1]$, and the range for E_{ce} is $p_o^{E_{ce}} \in [-1, +1]$. Values near 0.0 denote closeness to the main topic. The system uses a combination of weight, descriptor values and sequence in the sentence to mark some HBEs as *indicator words*.

A The Opposite HBE

From the WIH perspective, the contradiction is not directly processed. It remains to the reader the interpretation and solving of them. But when certain Spanish negative prefixes are detected the related HBE is detected as an *opposite HBE*. This kind of detection is used to bias the value of $p_o^{E_{ci}}$ related, and therefore to reflect some subjacent handling of words in the sentence and text. In these cases, the weight p_o is the multiplied by -1 . This value is combined with the rest of the HBE weights within the same E_{ci} (namely $p_o^{E_{ci}}$), resulting in a $p_o^{E_{ce}}$ weight. Table 1 shows a sample of prefixes and words considered this way. The main characteristics of this approach are:

- The opposite pattern list is mutable and perfectible. The original p_o value shares the same property.
- The patterns are related to the f_c active set. A change in the set could make a change in the actual list.
- The valuation is just a rough qualification of the original text words.

B Ambiguity in WIH

Several algorithms can be used to solve the ambiguity problem, most of them use Natural Language Processing (NLP) which is a hard task due the language expressions are related to writer's culture, education, geographical situation, etc. [8]. The WIH prototype [26] has many differences with NLP:

-In NLP it is required to process the entire document. WIH is not intended to process semantics but to extract certain features and words to represent it approximately.

-NLP works at a semantic level. WIH performs a simple morpho-syntactic processing. It does just an analysis in the neighbor of each word.

-NLP considers five main levels of information: lexical, morphologic, syntactic, semantic and pragmatic. WIH just makes use of some aspects from the first three levels.

-In NLP each level has to be robust enough to support upper abstraction levels. In WIH this is not necessarily true due to the simplicity of the process.

TABLE 1
SPANISH WORD WEIGHTS.

Opposite
No X
Sin X
desX
inhX
antiX
disX

The approach followed in the prototype does not intend to solve ambiguities but to detect and use them as a weighing bias of the word and sentence (through $p_i^{E_{ce}}$ and $p_o^{E_{ce}}$ respectively). Each time one of such words is detected, a special non-zero value is used for the corresponding HBE to express a vagueness degree of the related statement. Table 2 shows part of the Spanish words and prefixes detected. It is important to say that the list is mutable and related to the actual set of f_c as in the previous section's list.

V CASE STUDIES

The original test cases were extracted from web sites in Spanish but they were translated into English here. As the HBE describes many morpho-syntactic aspects of the related words, it has a complex structure. For simplicity they were replaced by the Spanish word from the site followed by its p_o value within parenthesis.

It is important to note that the system sometimes can decide to eliminate some words and not to represent them into a HBE. All the fields and data presented here are extracted automatically by the system.

A Contradiction

Three cases of Spanish contradictions were processed. For each one, the sentence is first presented in Spanish and translated in parenthesis. HBE for E_{ci} and E_{ce} are finally described with the originating words. All the web pages processed have the orchids as main topic.

1) Test case 1:

TEXT: "Las orquídeas vegetativamente son muy diferentes pero iguales." (Orchids are vegetatively very different but similar).

URL: orquidea.blogia.com temas – que-es-una-orquidea-.php
 E_{ci} WORDS: orquídeas (0.0) + vegetativamente (0.0) + son (0.0) + muy (0.7) + diferentes (0.0) + pero (0.0) + iguales (0.0)
 E_{ce} WORDS: orquid (0.0) + son (0.0)

As can be seen, the semantic contradiction is not taken into account and remains hidden in the original text. The system detects some important information (in fact a definition) and promotes two HBEs to E_{ce} level. At this time, the user that is looking for information at an E_{ce} level is notified there is a definition, but is not provided with the definition itself (the information content value at the E_{ce} is 0.0).

2) Test case 2:

TEXT: "Las estructuras reproductivas están fusionadas separadas." (Reproductive structures are merged separately).

URL: orquidea.blogia.com temas – que-es-una-orquidea-.php
 E_{ci} WORDS: estructuras (0.0) + reproductivas (0.0) + estan (0.0) + fusionadas (0.0) + separadas (0.0)
 E_{ce} WORDS: none

Table 2. Ambiguity patterns

Ambiguity
muy X
tan X
algo X
poco/a(s) X
mucho/a(s) X
bastante(s) X
escaso/a (s) X

The contradiction is not taken into account again. Even the system decides there is no important information in the sentence and does not promote any HBE.

3) *Test case 3:*

TEXT: "Tienen y no tienen dos tipos básicos de crecimiento simpodial.". (They have and don't have two main kinds of sympodial growth).

URL: [orquidea.blogia.com temas – que-es-una-orquidea-.php](http://orquidea.blogia.com/temas-que-es-una-orquidea-.php)

E_{ci}WORDS: Tienen (0.0) + no (-1.0) + tienen (0.0) + dos (0.0) + tipos (0.0) + básicos (0.0) + crecimiento (0.0) + simpodial (0.0)

E_{cc}WORDS: none

The contradiction is not taken into account again. The system does not find important information in the sentence and does not promote any HBE.

B *Ambiguity*

In this section three test cases for each kind of ambiguity were selected. The information provided is organized as in the previous section, but here indicator words of the E_{ci} are denoted with an asterisk to make explicit the ambiguity. An SAQ (Spanish Ambiguity Question) is included to make apparent the ambiguity in the Spanish language.

1) *Lexical ambiguity: Test case 1*

TEXT: "Ha despertado las más inimaginables pasiones en los hombres.". (It has inspired the most unimaginable passion in men/mankind).

AMBIGUITY QUESTION: whose passion is? Men/mankind?

URL: [es.wikipedia.org wiki Orchidaceae](http://es.wikipedia.org/wiki/Orchidaceae)

E_{ci}WORDS: ha (0.0) + despertado (0.0) + más (0.0) + inimaginables (-1.0) + pasiones* (0.0) + hombres(0.0)

E_{cc}WORDS: pasiones

The prototype finds the HBE related to the word pasiones to be the one that express the essence of the sentence and promotes it to E_{cc}. It is therefore denoting that the phrase is speaking something regarding orchid passion with lack of interest about whose passion it is. Note that the special combination of the morpho-syntactic characteristics with the closeness to p_o(inimaginables)⁰ results in the word pasiones to generate an indicator HBE.

2) *Lexical ambiguity: Test case 2*

TEXT: "se conocen plantas recolectadas a mediados del siglo pasado que todavía están creciendo y floreciendo saludables en muchas colecciones". (Last century collected plants are known to be still growing and flowering healthily in several collections).

SAQ: collection as a group or as part of collections stock?

URL: [es.wikipedia.org wiki Orchidaceae](http://es.wikipedia.org/wiki/Orchidaceae)

E_{ci}WORDS: conocen (0.0) + plantas (0.8) + recolectadas (0.0) + mediados (0.0) + del (0.0) + siglo* (0.0) + pasado (0.0) + todavía (0.0) + están (0.0) + creciendo(0.0) + floreciendo (0.0) + saludables (0.0)

E_{cc}WORDS: siglo

The prototype finds the HBE related to the word siglo to be the one that expresses the essence of the sentence and promotes it to E_{cc}. It expresses the sentence is describing something about century and orchids. Note that this indicator is found even with all p_o values being zero.

3) *Lexical ambiguity: Test case 3*

TEXT: "Las orquídeas son realmente las flores de lo superlativo ". (Orchids are really flowers of superlative).

SAQ: superlative as something big or cute?

URL: [es.wikipedia.org wiki Orchidaceae](http://es.wikipedia.org/wiki/Orchidaceae)

E_{ci}WORDS: orquídeas(0.0) + son (0.0) + realmente (0.0) + flores (0.0) + lo (0.0) + superlativo (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one (the original text describe main considerations about taking care of orchids and general description).

4) *Syntactic ambiguity: Test case 1*

TEXT: "¿Sabías que la orquídea Brassavola Digbyana es flor nacional de Honduras?" (Did you know Brassavola Digbyana orchid is the Honduras national flower).

SAQ: is it from Honduras or is the flower representing Honduras?

URL: [63.173.68.43 sites 7dias content.cfm id 276 PageName Insolit](http://63.173.68.43/sites/7dias/content.cfm?id=276&PageName=Insolit)

E_{ci}WORDS: sabías (0.0) + orquídea(0.0) + Brassavola (0.0) + Digbyana (0.0) + es (0.0) + flor (0.0) + nacional (0.0) + Honduras (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one. But it is important to say the original text described mainly the Brassavola's maintenance requirements.

5) *Syntactic ambiguity: Test case 2*

TEXT: "La parte de la flor que produce el polen." (The part of the flower that produces pollen).

SAQ: is it speaking of part of a plant or about a plant?

URL: [orquidea.blogia.com temas -que-es-una-orquidea-.php](http://orquidea.blogia.com/temas-que-es-una-orquidea-.php)

E_{ci}WORDS: parte(0.0) + flor (0.0) + produce (0.0) + polen (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one without studying the ambiguity.

6) *Syntactic ambiguity: Test case 3*

TEXT: "Las orquídeas son realmente las flores de lo superlativo ". (Orchids are really flowers of superlative).

SAQ: superlative as something big or cute?

URL: orquidea.blogia.com

E_{ci}WORDS: orquídeas(0.0) + son (0.0) + realmente (0.0) + flores (0.0) + lo (0.0) + superlativo (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one. It is important to say this case was extracted from a very small dictionary at the end of the page.

7) *Semantic ambiguity: Test case 1*

TEXT: "Se corta unos 3 cms por arriba y por abajo de la yema." (Perform a cut 3 cm upper and bellow the leaf bud).

SAQ: 3 cm bellow and 3 cm upper? Or 3 cm upper and just bellow the bud?

URL: orquidea.blogia.com

E_{ci}WORDS: se(0.0) + corta (0.0) + unos (0.0) + cms (0.0) + arriba (0.0) + abajo (0.0) + yema (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one.

8) *Semantic ambiguity: Test case 2*

TEXT: "Según he leído en algunas webs, se puede acelerar agregando al agua algunas hormonas, pero yo nunca las he usado, así que no opino al respecto." (According to some webs, it can be accelerated by adding some hormones, but I've never used them, so I don't give any opinion).

SAQ: did not use hormonas at all or did not use them for this?

URL: orquidea.blogia.com

E_{ci}WORDS: según(0.0) + he leído (0.0) + algunas (0.3) + webs* (0.0) + puede (0.0) + acelerar (0.0) + agregando (0.0) + agua (0.0) + algunas (0.3) + hormonas (0.0) + pero (0.0) + yo (0.0) + nunca (0.0) + he (0.0) + usado (0.0) + así (0.0) + no (-1.0) + opino* (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. Despite all the statement is not considered as part of the main topic, the system finds that there is a secondary topic regarding an opinion and other webs.

9) *Semantic ambiguity: Test case 3*

TEXT: "A cada yema le he quitado la especie de piel que la cubre en forma de triangulo para dejar al descubierto la yema con unas pinzas." (I have peeled off each leaf bud to leave uncovered them with tweezers).

SAQ: is the peel triangular? Has it been taken off with this cutting shape?

URL: orquidea.blogia.com

E_{ci}WORDS: cada (0.0) + yema (0.0) + he (0.0) + quitado (0.0) + especie* (0.0) + piel* (0.0) + forma* (0.0) + triángulo (0.0) + dejar (0.0) + descubierto (0.0) + yema (0.0) + unas (0.0) + pinzas (0.0) + piel (0.0) + cubre (0.0)

E_{cc}WORDS: especie, piel, forma

The prototype finds the words to express there is a mention of this topic but it does not find interesting to describe the technique details. The ambiguity is preserved within the E_{ci} to be solved by the reader.

10) *Anaphora: Test case 1*

TEXT: "Unas serán fértiles, otras estériles." (Some of them will be fertile, others will be sterile).

SAQ: who?

URL: orquidea.blogia.com

E_{ci}WORDS: unas(0.0) + serán (0.0) + fértiles (0.0) + otras (0.0) + estériles (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one.

11) *Anaphora: Test case 2*

TEXT: "Debajo de esta hay una pequeña "ramita"." (Below it there is a small branch).

SAQ: below what?

URL: orquidea.blogia.com

E_{ci}WORDS: ramita(0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one. In the original text it is a part of a more complete description on how to handle leaf buds.

12) *Anaphora: Test case 3*

TEXT: "Con ello conseguiremos que les llegue más luz." (With that we allow them to get more sunlight).

SAQ: who?

URL: orquidea.blogia.com

E_{ci}WORDS: con(0.0) + conseguiremos (0.0) + llegue (0.0) + más (0.0) + luz (0.0)

E_{cc}WORDS: none

The prototype doesn't find an indicator HBE. The weight values are also all 0.0. As a consequence all the statement is considered as an irrelevant one.

VI CONCLUSIONS

Some real cases handling contradictions and several ambiguities were presented. They were processed by WIH and handled as in any other case without being disturbed by the lack of enough context information. On the other hand, proposals like Classic [10] need to go further, with a detailed purpose clause analysis, and try to encode linguistic knowledge about actions as a way to provide some context to interpret new statements through a formal analysis, making the whole process too complex.

In spite of the fact WIH neither disambiguates nor eliminates contradiction, it is able to extract main words and represent the main sentence content into an E_{cc} level. In contrast, in systems like Asker [9] the contradictions and ambiguities are detected, collected and reorganized into structures that need to be constantly updated and used to overcome contradictions while performing text processing.

Other proposals such as Click-talp [24], perform a morpho-syntactic level annotation, while WIH does all its work skipping that step.

From all this, it can be said that p_o could be a simple tool to provide support to some NLP.

VII FUTURE WORK

There is a lot of work that remains yet: to tune and complete the list of p_o weights and to test them against real cases in order to validate results. Although this paper is a summary of the results on 100 web pages it is important also to perform a test with a higher number of pages. It will be very interesting to perform an extension of WIH to other languages as well.

ACKNOWLEDGMENT

The author gratefully acknowledges the contributions of Dr. J. Ale and Prof. M. Bosch for their work on the original version of this document.

REFERENCES

- [1] Bargalló M., Forgas E., Garriga C., Rubio A. Las lenguas de especialidad y su didáctica. Schnitzer Eds. Universitat Rovira i Virgili. Tarragona, cap. 1 (P. Schifko, Wirtschaftsuniversität Wien), pp. 21-29. 2001..
- [2] Alonso Pardo M. A. Interpretación tabular de autómatas para lenguajes de adjunción de árboles. PhD Theses. Spain. 2000.
- [3] José Padrón G. Notas Sobre Análisis Del Lenguaje. Cap. 1: modelo GENERAL de los lenguajes. Maracaibo. La Universidad del Zulia - Doctorado en ciencias Humanas. 1997.
- [4] Butt M. XLE: Grammar Development Platform. Proc of ES SLLI-2000. 2000.

- [5] Malheiro B., Oliveira E. Improving Assumption-based Distributed Belief Revision. SCAI 95. pp 41-50. 1995.
- [6] Gómez-Gauchia H., Díaz Aguado B., González Calero P. Two-Layered approach to Knowledge representation using conceptual Maps and Description Logics. Proc. Of CMC2004. 2004.
- [7] Gómez-Gauchia H., Díaz Aguado B., González Calero P. Ontology-Driven Development of Conversational CBR Systems. Proc. Of ECCBR. 2006.
- [8] Anderson M., Bathia M. Handling Uncertainty with Active Logic. Proc. of AAAI. 2001.
- [9] Bobrow D. B., Condoravdi C., de Paiva V., Karttunen L., King T. H., Nairn R., Price L., Zaenen A. Precision-focused Textual Inference. Proc. of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. 2007.
- [10] Di Eugenio B. Understanding Natural Language instructions: A Computational Approach to Purpose Clauses. PhD thesis. 1993.
- [11] Crouch D., et. al. Entailment, Intentionality and Text Understanding. Proc of HLT-NAACL. Vol 9, pp. 38-45. 2003.
- [12] Bernard G. Compréhension de texte, opérations linguistiques, linguistique textuelle, reference. 1990.
- [13] Bernard G., Feat J., San Gines P., Sabido V. Comprension de textos: elementos de modelizacion. Caminos del texto (Actes des colloques de Madrid et Grenade). Universidad de Granada. 1995.
- [14] Ruch P., Gaudinat A. Comparing corpora and lexical ambiguity. Workshop on Comparing Corpora, Proc. of ACL, Hong-Kong. 2000.
- [15] Kennedy C., Boguraev B. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. Proc. of the 16th conference on Computational linguistics - Volume 1. Copenhagen, Denmark. pp. 113 - 118.
- [16] Montemagni S., et.al. Resolving syntactic ambiguities with lexico-semantic patterns: an analogy-based approach. Proc. of the 16th conference on Computational linguistics - Vol 1. Copenhagen, Denmark. pp. 376 - 381. 1996.
- [17] Levinger M., Ornan U. Learning Morpho-lexical Probabilities from an Untagged Corpus with an Application to Hebrew. Computational Linguistics. Volume 21, Issue 3. pp. 383 - 404. 1995.
- [18] Prószyński G., Naszódi M., Kis B. Recognition Assistance: Treating Errors in Texts Acquired from Various Recognition Processes. Proc. of the 19th international conference on Computational linguistics - Vol 2. pp 1 - 5. 2002.
- [19] Langer H. Syntactic Normalization of Spontaneous Speech. Proc. of the 13th conference on Computational linguistics - Vol 3. Helsinki, Finland. pp 180 - 183. 1990.
- [20] Rosario B., Hearst M. Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy. Proc. of Conference on Empirical Methods in Natural Language Processing. p. 82-90. 2001.
- [21] Rosario B., Hearst M.A., Fillmore C. The Descent of Hierarchy, and Selection in Relational Semantics. Meeting of the Association for Computational Linguistics (ACL). 2002.
- [22] Cha J., Lee G. Structural disambiguation of morpho-syntactic categorical parsing for Korean. Proc. of the 18th conference on Computational linguistics - Vol 2. Saarbrücken, Germany. pp. 1002 - 1006. 2000.
- [23] Kinyon A. Hypertags. International Proc. of 18th conference on Computational linguistics - Vol 2. Saarbrücken, Germany. pp. 446 - 452. 2000.
- [24] Civit M., Martí M.A. Estándares de Anotación Morfosintáctica para el español. Taller de Herramientas y Recursos Lingüísticos para el Español y el Portugués. Tonantzintla, México. 2004.
- [25] Martínez Fernández P., García Serrano A.M. Interacción persona-web empleando recursos lingüísticos. Revista Iberoamericana de Inteligencia Artificial. N16. pp 55 - 65. 2002.
- [26] López De Luise M. D. A Morphosyntactical Complementary Structure for Searching and Browsing . Proc. Of SCSS 2005. Springer. pp. 283 - 290. 2005.