Automatic Retrieval of Definitions in Texts, In Accordance with a General Linguistic Ontology

Charles Teissedre
Sorbonne University
28 rue Serpente
75006 Paris France
charles.teissedre@gmail.com

Brahim Djioua
LaLICC - Sorbonne University
28 rue Serpente
75006 Paris France
bdjioua@gmail.com

Jean-Pierre Desclés
LaLICC - Sorbonne University
28 rue Serpente
75006 Paris France
Jean-pierre.descles@paris4.sorbonne.fr

Abstract

A semantics of definition category and its sub-categories used in texts, organized in a semantic map, requires a more complex structuring of the domain underlying the meaning representations than is commonly assumed. This paper proposes a three-layer ontology in which the notion of definition takes part and indicates how it can be used in Information Retrieval. The first part describes an automatic process to annotate definitions based on linguistic knowledge, in accordance with a general linguistic ontology, and the second part shows a practical use of its semantic and discourse organizations in retrieving information through the Web.

General introduction

A large variety of information processing applications deals with natural language texts. Many of these applications require extracting and processing the meanings of texts, in addition to processing their morpho-syntactic forms. In order to extract meanings from texts and manipulate them, a natural language processing system must have a significant amount of knowledge about the organization of semantic and discourse notions. We can also observe that the focus of modern information systems is moving from ”data processing” and ”concept processing” towards ”relation between concept processing”, which means that the basic unit of processing is less and less an atomic piece of data and tends to be some more general semantic and discourse organization of texts. We already made a process for automatic building of domain ontology with semantic and discourse relations related to a linguistic general ontology and terminology for a specific domain [Le Priol et ali., 07].

The debate about general and domain ontology is about which approach to domain categorization is ’best’ ? [Poesio, 05]:

- Designing a clean, elegant ontology with a clear semantic based on sound philosophical principles and scientific evidence,
- Relying on evidence from psychology and corpora, and on machine learning techniques, to acquire - automatically, as far as possible - a domain structure that in most cases will be rather messy.

This paper describes the semantic and discourse organization of the linguistic notion of ”definition” in accordance with a general linguistic ontology and its use in Information Retrieval. After a presentation of some general ontologies, we describe the notion of definition in texts, before showing a practical use in information retrieval through the Web.

General ontologies versus Domain ontologies

According to Wikipedia a general ontology is defined as following:

> In information science, an upper ontology (top-level ontology, or foundation ontology) is an attempt to create an ontology which describes very general concepts that are the same across all domains...[1]

The goal of this is to construct broad accessible ontologies resulting from these Upper-Ontologies. An Upper-Ontology is often presented in the form of a hierarchy of entities and of their associated rules (theorems and constraints) which try not to hold account of a particular issue in specific domain. It appears increasingly that more than the domain ontologies, general ontologies are economically relevant.

[1] http://en.wikipedia.org/wiki/Upper_ontology_\%28computer_science\%29)

Recognizing the need for large domain-independent ontologies, diverse groups of collaborators from the fields of engineering, philosophy and information sciences have come to work together to Upper-Ontologies like (i) CyC, an artificial intelligence project that attempts to assemble a comprehensive ontology and database of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning, (ii) Suggested Upper Merged Ontology or SUMO, an upper ontology that stands for a foundation ontology for a variety of computer information processing systems. It is one candidate for the ”standard upper ontology” that IEEE working group 1600.1 is working on; (iii) DOLCE designed by N. Guarino and his group which is based on a fundamental distinction between enduring and perduring entities (between what philosophers usually call continuants and occurrents); (iv) GOLD, a linguistic ontology.

It gives a formalized account of the most basic categories and relations used in the scientific description of human language. Other sources for inspiration are the lexical and ontological projects that are being developed for computational linguistics, Natural Language Processing and the Semantic Web. Pustejovsky and his team propose a general framework for the acquisition of semantic relations from corpora, guided by theoretical lexicon principles according to the Generative Lexicon model. The principal task in this work is to acquire qualia structures from corpora.

At present, no Upper-Ontology description became de facto a standard, even if some of them alike NIST[2] try to define the outlines of a standardization with, for instance, the proposition of a standard for the specific domains like PSL (Process Specification Language) which is a general ontology for describing manufacturing processes that supports automated reasoning.

The general ontology used in our work to describe the organization of the notion of *definition* and its use in texts is organized in three layers [Desclés, 2007]:

1. **domain ontologies** with instantiation of linguistic relations, discursive (quotation, causality, definition) and semantics (whole-part, spatial movement, agentivity) projected on terminologies of a domain.
2. the intermediate level presents a certain number of **discursive and semantic maps** organizing different linguistic categorizations.
3. the high level describes the **representation of various categorizations with Semantico-Cognitive Schemes** in accordance with the linguistic model of the Applicative and Cognitive Grammar [Desclés, 1990].

The second-level concepts of our ontologies organization are structured in semantic maps, that is to say networks of concepts. The instances of these concepts are linguistic markers. The relations between concepts, in a semantic map, are relations such as ingredience, whole-part, inclusion, subclass-of, etc. Thus, the point of view of definition is associated with a semantic map presented further in this article.

The concepts of semantic maps can be analyzed with the aid of semantic notions described in the third layer, based on the semantic and cognitive concepts of the Applicative and Cognitive Grammar. For instance, if a second-level-concept contains spatial and temporal relations then we must use, to describe it, more general concepts (third-level concepts) such as, in time domain: ”event”, ”state”, ”process”, ”resulting state”, ”consequence state”, ”uttering process” , ”concomitance”, ”non concomitance”, ”temporal reference frame” ; or, in space domain: ”place”, ”interior of a place”, ”exterior of a place”, ”boundary of a place”, ”closure of a place”, ”movement in space”, ”oriented movement”, ”movement with teleonomy”, ”intermediate place in a movement”. Other general concepts must be also defined with precision, for instance: ”agent who controls a movement”, ”patient”, ”instrument”, ”localizer”, ”source and target” [Desclés, 2007].

Definition as a text mining point of view

A user’s search for relevant information proceeds by guided readings which give preferential processing to certain textual segments (sentences or paragraphs). The aim of this hypothesis is to reproduce ”what naturally makes a human reader”, who underlines certain segments related to a particular point of view which focuses his/her attention. There are several points of view for text exploration; they correspond to various focusing on more specific research of information. Indeed, such a user could be interested, while exploring many texts (specialized encyclopaedias, handbooks, articles), in the definitions of a concept (for example ”social class” in sociology, ”inflation” in economy, ”grapheme” in linguistics, etc).

The aim of these points of view for text mining is to focus reading and possibly annotate textual segments, which corresponds to a research guided in order to extract information from them. Each point of view is explicitly indicated by identifiable linguistic markers in texts. Our hypothesis is that semantic relations leave some discursive traces in textual documents. We use cognitive principles which are based upon the linguistic marks found in texts, in the organizing discursive relations. For instance, we use the following linguistic marks *we define ...as, use ... to*

[2] National Institute for Standards and Technology (http:// www.nist.gov)

denote or *which means*, *...is, by definition,...* to extract defining relations.

Definition in texts

The aim of this study is both to circumscribe and analyze, on the very surface of the language, the linguistic marks of definition : this theoretic and linguistic approach will lead us to an applicative approach, trying to extract definitory utterances on a linguistic-based ground.

Definition, in texts, can be seen as a relation between a <u>definiendum</u> (the defined entity) and a <u>definiens</u>, the defining proposition, which aim is to delimit the definiendum meaning (its essence). In texts, some definitions of a concept appear to be very general and are presented as having an universal scope, such as, for instance, definitory properties of numbers or of time. On the other hand, other expressions of definitions are linked with the stances of an author or with the description of a specific domain, such as, for instance, the notion of number as children apprehend it from the specific point of view of psychology or the notion of number apprehended by ethnologists. Indeed, some notions are not always considered and defined the same way: for instance, social classes, in sociology, are defined very differently from an author (or a theory) to another.

From a general point of view, in philosophy, in logic and in linguistics, a concept ‘f’ is characterized by its intension, i.e. the class of concepts that this concept ‘f’ comprises or entails. The extension of a concept is the class of all its instances. In the LDO approach (the Logic of Determination of Objects), presented earlier in FLAIRS by [Desclés and Pascu, 2005], the concepts of the intension are only inherited by the typical instances, while all instances, both typical and atypical instances, necessarily inherit of the concepts of the essence, the essence being a part of the intension. The definition of a concept should generate the essence and should characterize all instances of the extension (typical and atypical). However, for a complex concept of a domain, we are not always able to produce proper definitions, but only to give partial definitory relations or facets of definitions. For instance, there is no definition of mankind that reaches a consensus: in texts, we can generally only find facets of its definition, such as “man is a reasonable biped”. This characterization is not a plain definition of mankind, as some atypical men exist that are not bipeds and as children may not be considered as reasonable, though they should both be categorized as human.

Retrieving definitions in texts enables to compare, to differentiate and to confront authors stances and theories, as a same concept can generate different definitions or facets of definition.

Our linguistic approach to achieve the automatic retrieval of definitions in texts relies on the Contextual Exploration method [Desclés, 91, 97, 06], which consists in identifying textual segments that correspond to a semantic ”point of view”. In our research of linguistic marks with which the definition point of view is expressed in texts, the method introduces a hierarchy between strong indicators of definition and linguistic clues found in their context : those clues (any kind of linguistic units : lexical, grammatical or typographic) help to lift the ambiguity on the semantic value of an indicator, so that it gets possible to infer that the segment in which the indicator appears is a definition.

Intuitively, we can gather several strong indicators of definition, such as ”to define”, ”to mean”, ”to denote” : but even those obvious verbal indicators are highly polysemous and require an exploration of their context to determinate the value they held. An example of a non definitory utterance can illustrate this aspect :

(1) The lawyers proceed <u>to define</u> taxes, tolls, and exactions of various kinds to be imposed on trade.

This means, that after collecting strong indicators (mostly verbs, but not exclusively), it is necessary to bring out regularities in their context of apparition that facilitate semantic disambiguation. Different features appear to be very regular : for instance, if the expression ”the word xxx” (or any approaching expression) appears in the left context of the verbal indicators we’ve mentioned, the sentence is likely to be a definition :

(2) Plato used <u>the word</u> aeon <u>to denote</u> the eternal world of ideas, which he conceived was ” behind ” the perceived world (...).
(3) <u>The verb</u> krstiti in Croatian <u>means</u> ” to baptize ”.
(4) <u>The term</u> abstract algebra now <u>refers to</u> the study of all algebraic structures, as distinct from the elementary algebra ordinarily taught to children (...).

Among the important and frequent features, several interpolated clauses help to raise the ambiguity of an indicator : these clauses are often traces left by the definitory work. Among those, we can count rewording clauses and parenthesis (”*,also called,*”, ”*(or…)*”, etc.) and clauses related to signification (”*, in its literal meaning,*”, ”*, in a metaphorical sense,*”, ”*, strictly speaking,*”, ”*by extension,*”, ”*in a philosophical terminology,*”, etc.). These clauses help to disambiguate many identification utterances : indeed, identification utterances (introduced by ”is a” or ”is the”) may not be confused with definitory utterances, though they may be frequent ingredients of definition. In some case, though, some identification utterances may stand for definitions - but definitions in which the definitory work is silent.

Beyond these, there are many others linguistic clues we haven't mentioned : they all are organized in some way around the indicators, on their left or right context. A linguistic work progressively brings out these organizations and formalizes them in Contextual Exploration rules that express discursive and semantic configurations. The linguistic analysis our work has carried out shows that definition can be organized in the following semantic map (figure 1) :

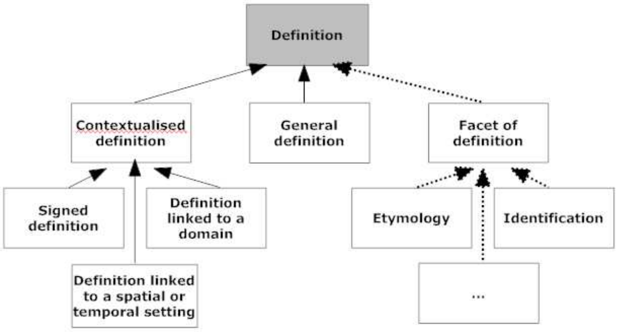

Definition
Contextualised definition
General definition
Facet of definition
Signed definition
Definition linked to a domain
Etymology
Identification
Definition linked to a spatial or temporal setting
...

Fig. 1: semantic map of definition

The semantic map shows the different values that a definition can held: a definition can either be general (universal scope) or, on the contrary, be contextualized. In this case, the context of validity of the definition can be of different kinds : the definition can be signed (by an author); it can also be linked to a specific domain or contextualized in time ("*In the XIXth century, X was defined as…*"). As shown, a facet of definition is not a proper definition: it is an ingredient of definition. The link with the general class of definition is not a link of subclass, but a link of ingredience.

Automatic semantic annotation and indexation of definitions

We can find several attempts to use linguistic tools in the semi-automatic process of populating domain ontologies from texts [Mädche and Staab, 2000] [Cimiano, 2006] by (i) using collocations that typically reveal a strong but unknown relation between words; (ii) using Syntactic Dependencies, in particular the dependencies between a verb and its arguments.; (iii) using lexico-syntactic patterns defined originally by Hearst for relations such as part-of, cause, …; and (iv) learning Qualia Structures [Pustejovsky, 1991] thus a Qualia structure describes a fixed set of relations which every object possesses. Some of these relations are the part-whole and subclass-of relations.

In our approach, the semantic of each relation between concepts in domain ontologies can be analyzed by a lambda-expression with intrinsic mathematic properties alike functional type [Le Priol and al., 2007]. In order to populate ontologies, the retrieval of definitions and facets of definitions in texts can be an aid to the construction of the relations between concepts, as well as of the relation between concepts and instances.

Automatic annotation processing

As for us, the methodology used by the general automatic annotation engine EXCOM [Djioua & al. 06], called Contextual Exploration [Desclés & al. 91, Desclés 06], describes the discursive organization of texts exclusively using linguistic knowledge present in texts. Linguistic knowledge is structured in the form of lists of linguistic markers and declarative rules fixing the way to explore the context of linguistic indicators retrieved in texts. The constitution of this linguistic knowledge is independent of a particular domain. Domain knowledge describes the concepts and sub-concepts of a subject domain with their relationships. Contextual knowledge concerns communicative knowledge as a discursive organization, which deals with the preferences and needs of those who research information in texts. Linguistic rules define different strategies for identifying and semantically annotating textual segments. Some of these rules use lists of simple patterns coded as regular expressions, others need to identify structures like titles, sections, paragraphs and sentences for extraction purposes.

EXCOM engine explores the semantic and discursive organizations of text, in order to annotate pertinent segments with semantic tags, considering a given 'point of view' : in the present study, the 'point of view' chosen to explore texts is the definition angle. The tags with which texts are enriched are the values given by the semantic map of the exploration point of view. Indeed, EXCOM is designed in accordance with the Contextual Exploration method. The knowledge implemented in the engine only consists in a list of linguistic indicators or clues and in rules that refer to those lists. The EXCOM rules are gathered in an XML document. Each rule declares a condition to launch a Contextual Exploration : if the condition is fulfilled (eg. if an indicator is found), declared actions are performed (eg. looking forward or backward for the presence of other linguistic clues). The figure 2 shows an example of those rules.

Designing Excom rules for the definitory point of view meant encoding the discursive organizations brought out by the linguistic analysis we briefly described. Lists of indicators were collected and the layout of linguistic clues around the indicators have then been described in Contextual Exploration rules. The major characteristic of this work in comparison with other points of view that have been implemented in EXCOM (quotation, causality, etc.) is that the number of indicators is rather small, while the variety of discursive layouts around these indicators is, on the contrary, rather large. This explains that, for a single indicator, numerous rules were implemented.

```
<!-- the word...means/denotes... -->
<rule rule_name="Word+DefinitoryVerbs"
        task= "RelationsBetweenConcepts"
        point_of_view= "Definitions_and_Definitory_Facets"
        type="EC">
<conditions>
<indicator research_range="sentence"  type="list"
              value="denotes|means|refersTo" />
<clue context="left" research_range ="." type=" list "
        value ="WordAndSynonyms" />
</conditions>
<actions>
<annotation annotation="Definition" range="same"
                type="add_attribute" />
</actions>
</rule>
```

Fig.2: An example of Contextual Exploration rule

Automatic indexation of definition utterances

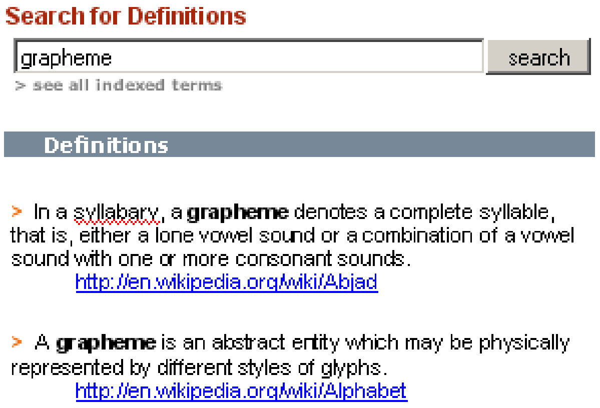
The general process of indexing uses a multilevel structure composed by textual segments (such as titles, sentences, paragraphs, sections, etc) and discourse and semantic annotations (such as causality or part-whole relations). This organization makes explicit the relationship between the initial document, the constitutive textual segments (sentence, paragraph and section title), the discursive and semantic annotations, and the extracted terms that compose the textual segments. A document is considered, in this model, as a set of textual segments identified automatically by EXCOM together with its discursive and semantic organizations expressed by the author. Each textual segment is associated with several important pieces of information, namely:

- a set of semantic annotations (discursive marks such as definition). Each textual segment identified as relevant for a document can be associated to a set of discursive and semantic annotations according to the semantic points of view used in the annotation engine EXCOM. Then, the same textual segment can be chosen by the search engine MOCXE as an answer to a query concerning a definition.
- the document's URI that ensures its unique identification on the Internet.
- the document's title in order to have a better readability of the answer.
- the terms found in the full-text content for a relevant answer to users. The procedure used to index textual segments is the same as the methodology used by Salton.

The indexing engine MOCXE generates an index of textual documents by using the output of the annotation engine EXCOM. The index gives the possibility to retrieve information according to discourse and semantic relations (definition). The queries are formulated by using semantic relations that are defined inside a point of view for text mining; the answers are given in the form of ”annotated textual segments”. For indexing process we use techniques already available in open-source software for the search engine Lucene/Nutch architectures (www.apache.org).

Results

The implementation of EXCOM-MOCXE engines for definition retrieval have been tested on encyclopaedic corpora, as this type of texts presents a great amount of definitions. These tests have been realized on French resources, mostly on randomly chosen articles of Wikipedia. Definitions were extracted and categorized in accordance with the semantic map, even though the analysis has not always been able to distinguish subtly inside the big categories : the recognized categories are mainly ”definition”, opposed to ”facet of definition”, and, within the ”facets of definition”, the category of ”etymology”. The figure 3 shows an example of the outputs given by EXCOM-MOCXE engines for a query on “grapheme”.



Search for Definitions

grapheme

> In a syllabary, a **grapheme** denotes a complete syllable, that is, either a lone vowel sound or a combination of a vowel sound with one or more consonant sounds.
http://en.wikipedia.org/wiki/Abjad

> A **grapheme** is an abstract entity which may be physically represented by different styles of glyphs.
http://en.wikipedia.org/wiki/Alphabet

Fig.3: an example of EXCOM-MOCXE engines output

The tests that have been carried out on the French resources showed that 89.3% of the annotated segments are correctly annotated. Three hundred annotations were randomly extracted from the index of defined entities and validated by a human annotator. The protocol to conduct these tests consisted in attributing scores to the annotated textual segment linked to each defined entity : if an annotated utterance was correctly annotated, the score was 1, if not, the score was 0 or half a point if the defined entity was not perfectly segmented (for instance, the segment “amoeba or ameba” was considered as a plain entity, while it should have been divided in two). A wider validation process will be carried out on some more eclectic corpora, as the EXCOM rules are designed to analyze any textual document, regardless for its subject or type.

Another test has been played out through a comparison with a service of Google, specifically dedicated to definition retrieval. The aim of this second experiment is to illustrate the interest of semantic and discursive annotation and indexation.

A Comparison with ”Google Define”

The comparison with the ”Google Define” service has been realized through a short but instructive test in French that intended to show the relevance of a full-text semantic analysis. ”Google Define” seeks for documents that are shaped like glossaries : it does not proceed to any full-text analysis, but mostly searches for HTML markers and descriptors of glossaries, coupled with statistical computing. In order to point out the difference of such an approach with EXCOM-MOCXE engines, we developed a test concerning three concepts which we tried to define : *sustainable development*, *ontology* and *bird flu*. To collect definitions on those concepts, we gathered the first hundred documents answered by Google to an ordinary query on each of those terms. The three hundred documents collected were then analyzed by EXCOM, launched with the definitory point of view. MOCXE engine then indexed the definitions annotated in the documents. After this chain of processes, it became possible to make requests on the indexed database.

For the same queries, “Google Define” returned 28 utterances, 24 of them being relevant (corresponding to definitions) : 1 definition for “bird flu” (in French, “grippe aviaire”), 10 for “ontology” (“ontologie”), 13 for “sustainable development” (“développement durable”) and 4 irrelevant answers. EXCOM-MOCXE engines brought out 25 definitions on 26 utterances indexed : 7 definitions for “bird flu”, 13 for “ontology”, 5 for “sustainable development” and 1 irrelevant answer.

Those results show that on a tiny portion of Google database, our semantic and discursive approach makes it possible to extract knowledge that Google Define does not retrieve, as all the definitions that EXCOM-MOCXE engines extracted were different from the one given by Google Define. Indeed, Google Define does not really proceed to a text analysis, but mostly seeks for glossaries. Furthermore, in EXCOM text-mining, the utterances extracted are categorized and differentiated. It is possible for the user to specify which precise category of definition he is interested in. This very quick test illustrates the originality of a semantic indexation that does not process key-words only, but full textual segments to which semantic values are bound.

Conclusion

The results obtained for a couple of queries show what benefit brings a semantic and discursive approach in IR : with EXCOM-MOCXE engines it is possible to extract definitions from any kind of texts, without restriction on theme, genre or type ; the user is given access directly to the information he is interested in, as it is textual segments that are returned and not lists of Web sites that he must look down on.

One main ambition of this project, developed by Lalicc laboratory, is to clarify, by categorizing the structures of cognitive actions and language, what means to seek for information. How do we proceed to convert information into knowledge? What piece of information is relevant : data, concepts, relations between concepts? How is knowledge expressed in texts? How could this knowledge be categorized? The present study of the definitory point of view is a part of a wider project which aim is to facilitate access to knowledge.

References

Cimiano, P. eds. 2006. *Ontology Learning and Population from Text*, Springer

Desclés, J.-P. eds. 1990. *Langages Applicatifs, Langues Naturelles et Cognition*, Hermès, Paris, 1990.

Desclés J-P., Jouis C., Oh H-G., Reppert D. eds. 1991. *Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l’indicatif dans un texte*, in Knowledge modeling and expertise transfert, Eds D. Hrin-Aime, R. Dieng, J-P. Regourd, J-P. Angoujard, 371-400. Calif : IOS Press.

Desclés, J-P., Pascu, A. eds 2005. *Logic of Determination of Objects : the meaning of variable in quantification*, FLAIRS 2005, Florida.

Desclés, J.-P. eds 2007. *Ontologies, Semantic Maps, and Cognitive Scheme*, FLAIRS 2007, Florida.

Djioua B. et al. eds 2006. *EXCOM: an automatic annotation engine for semantic information*, FLAIRS 2006, Florida, May 11-13.

Djioua B., Desclés J.-P. eds 2007. *Indexing Documents by Discourse and Semantic Contents from Automatic Annotations of Texts*, FLAIRS 2007, Special Track ”Automatic Annotation and Information Retrieval : New Perspectives”, Key West, Florida, May 9-11.

Le Priol F., Djioua B., Garcia D. eds 2007. *Automatic Annotation of Discourse and Semantic Relations supplemented by Terminology Extraction for Domain Ontology Building and Information Retrieval*, FLAIRS 2007, Special Track ”Automatic Annotation and Information Retrieval : New Perspectives”, Key West, Floride, May 9-11.

Mädche, A. and Staab, S. eds 2001. *Ontology Learning for the semantic web IEEE Intelligent Systems*, 16(2):72:79

Poesio M., (2005), Domain modelling and NLP: Formal ontologies? Lexica? Or a bit of both? Applied Ontology Vol 1. 27-33.

Pustejovsky J., eds 1991. *The generative lexicon, Computational Linguistics*, 17(4):209-441.