

*Hélène Włodarczyk*

e-mail: helene.wlodarczyk@gmail.com

*André Włodarczyk*

e-mail: wlodarczyk.andre@gmail.com

Centre for Theoretical and Applied Linguistics (CELTA)

Paris Sorbonne University (1999–2014)

## The Interactive Method for Language Science and Some Salient Results

DOI: <http://dx.doi.org/10.12775/ZN.2019.025>

**Abstract.** The use of information technology in linguistic research gave rise in the 1950s to what is known as *Natural Language Processing*, but that framework was created without paying due attention to the need for logical reconstruction of linguistic concepts which were borrowed directly from barely (or even not at all) formalised structural linguistics. *The Computer-aided Acquisition of Semantic Knowledge* project (CASK) based on the Knowledge Discovery in Databases technology (KDD) enabled us to interact with computers while gathering and improving our knowledge about languages.

Thus, with the help of data mining tools, as a result of revisiting two sorts of generally admitted linguistic theories (the *Predicate Argument Structure* theory and *Information Structure* theory), we succeeded in improving these local linguistic approaches by proposing to unify the *Associative Semantics* (AS) theory (in which we introduced the concept of **ortho-information**) with the *Meta-Informative Centering* (MIC) theory (in which we described the **meta-informative** layer of natural languages). The resulting Distributed Grammar (DG) program (sketched out in this paper) treats, in addition to the above types of information, the third one, **para-information** (concerning identity and likeness with respect to context and language ontology) which – despite many studies – had no uniform theoretical background in general linguistics. This DG program aims to lay the foundations for creating the theoretical background of *Conceptual Linguistics*.

**Keywords:** epistemology; language philosophy; semiotics; linguistics; information; data mining; formal concepts; interactive method

### 1. Introduction

The interactive method as applied to linguistics has become a central procedure within research in language science. This was due to the application of algorithms which have been developed mostly in data mining technology to enhance knowledge extraction from data. Those methods elaborated in computer science gave rise to new procedures for interdisciplinary research via innovative *interaction* with a computer, aiming at the development of an *integral* theory of language. Applying

interactive methods<sup>1</sup> to our MIC theory,<sup>2</sup> we could develop the Distributed Grammar (DG) program whose philosophy, in turn, made it possible to lay the foundations of a new trend in the realm of *conceptual* theories of cognition.

But what are distributed systems? As they contain elements whose characteristics are irrelevant for the whole system, distributed information systems are neither (1) “families of formal contexts” in the sense of the Formal Concept Analysis (Wille 1982) nor (2) “information systems” in the sense of the Rough Set Theory (Pawlak 1991). Nevertheless, as an example of a distributed system we might quote the so-called “Resemblance Family” (Wittgenstein 1953). It is as though any (different) absurd tolerating “information systems” were combined into a single binary system by virtue of the fact that all their objects have something in common. In natural languages, this “something in common” does not even have to be semantic at all. Here, it is sufficient that, as a result of the evolution of a given language, certain semantic categories are combined together into units, either morphological or syntactic. It is probably worth mentioning that, surprisingly, some linguists use this unexpected peculiarity of natural languages, pushing their theoretical efforts so far as to consider that the resulting combined categories are hybrid. Such theories can hardly be thought of as scientific.

In this paper, we present (1) some methods and tools of data science useful for experimental interactive linguistics and (2) the impact of such experiments on the formation of innovative language theoretical solutions.

## 2. Some methods and tools of interactive linguistics

At the beginning of the third millennium, more and more linguists are showing interest in using ideas, methods and tools elaborated in computational intelligence for their research, aiming at building or logically reconstructing<sup>3</sup> (enhancing, integrating and formalising) structural theories of language in order to conceive meta-theoretical mathematical foundations. *Interactive Linguistics* (henceforth IL) makes use of computer tools to build models of linguistic theories following the

---

<sup>1</sup> Interactive Linguistics has emerged from an original line of research into language studies elaborated at CELTA (Centre de Linguistique Théorique et Appliquée), Sorbonne University (Paris), at the beginning of the 21st century (1999–2014). Among other international cooperation projects, CELTA participated in two French-Polish scientific projects: (1) CASK, a PAI Polonium research project 2006–2007 with Jagiellonian University and (2) Interactive Linguistics, an invited session, organised by Wrocław University of Technology, during the KES International Conference in 2020.

<sup>2</sup> This theory focuses mainly on the pragmatic nature of predication and explains how attention shapes utterances (Włodarczyk A, Włodarczyk H. 2013, 2016, 2019), see section 4.2 below.

<sup>3</sup> In Poland, it was Jerzy Pogonowski (1981, 1993), a mathematician, who – as a precursor of this trend – tried to initiate this direction of research by reconstructing the theory of a Danish functionalist structuralist linguist, Louis Hjelmslev (1899–1965).

example of empirical sciences (Stacewicz, Włodarczyk 2010). Model-based computing has already begun to bridge the gap between linguistic science and the formal method of experimental sciences (Włodarczyk A. 2015).

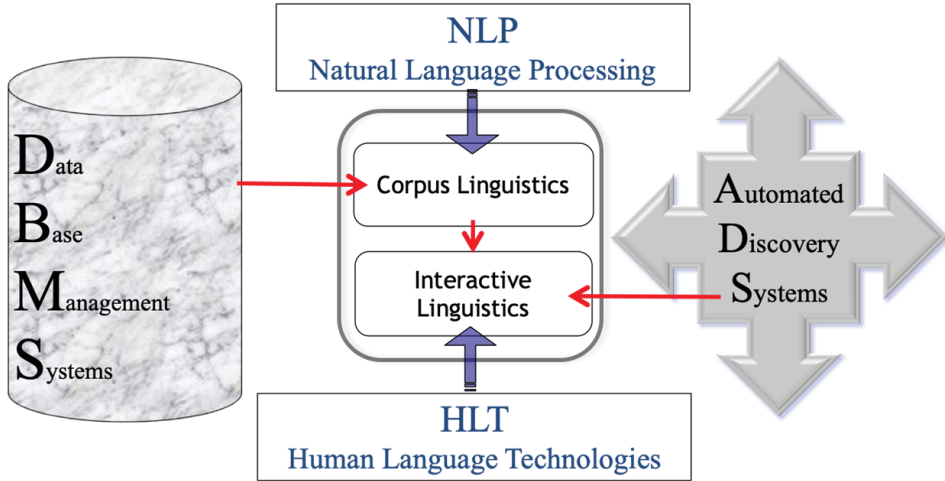


Fig. 1. Interactive Linguistics and Corpus Linguistics

However, *IL* differs from *Corpus Linguistics* (Fig. 1) as developed under the influence of Natural Language Processing (NLP) which consisted essentially in parsing algorithms for analysing natural language and relied on syntax and phonology with poor insight into semantics and pragmatics. Today, as shown in Fig. 2, *Corpus Linguistics* provides methods and tools for “in-large” research (text mining) while *IL* is concerned with “in-depth” research (data mining) including semantics and the specific nature of information conveyed by natural languages. *IL* aims at providing the best research standards for linguistic science while following the prominent results of building the semantic web in the field of information technology (IT). Its methods include both initial theoretical assumptions and interdisciplinary meta-theoretical knowledge, involving the scientific cooperation of linguists, logicians, psycho-neurologists and information engineers.

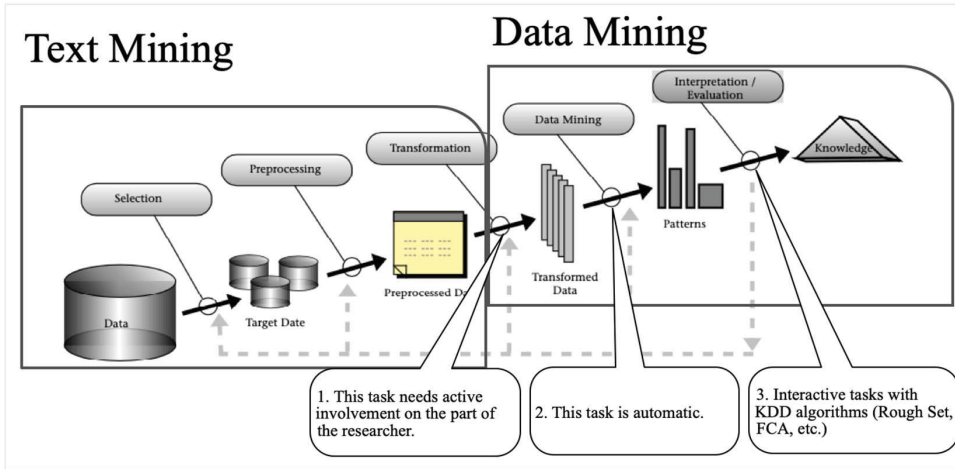


Fig. 2. Sketch by Fayyad U. et al. (1996) with our add-ons (subdivisions and balloons)

We consider that the task of text mining within Corpus Linguistics is to collect *data*. However, such data constitute only chunks of knowledge which, in turn, need to be transformed and annotated by the linguist in order to be gathered in datasets for further treatment by the interactive application of *data mining* algorithms (see sections 2.1 and 2.2 below).

## 2.1. Building data bases for linguistic research

A tool of interactive (computer-aided) discovery of ontology-based definitions of feature structures was designed at Sorbonne University especially for linguistic research (cf. CASK<sup>4</sup> project). This software called “Semana”<sup>5</sup> integrated a dynamic database builder with powerful functionalities of symbolic and statistical data mining tools (Fig. 3).

<sup>4</sup> The acronym CASK for computer-aided acquisition of semantic knowledge (CASK) was used in the early phase of our research (2004–2008).

<sup>5</sup> The *Semana* software was designed for CELTA by André Włodarczyk (2007) and Georges Sauvet (2008).

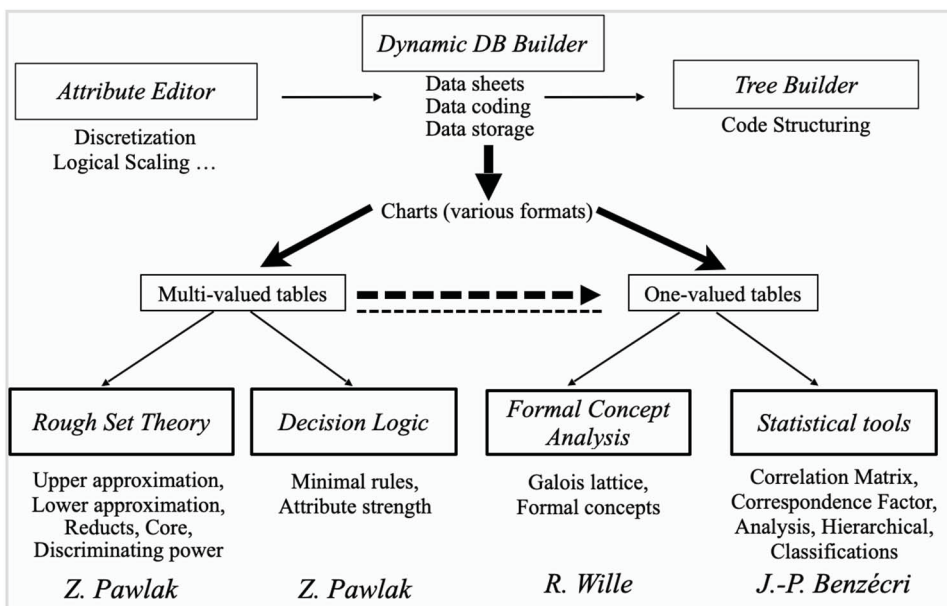


Fig. 3. The architecture of SEMANA

The dynamic DB Builder, due to its interactivity and interoperability, is a constructive environment with facilities for restructuring data. It is always possible to modify features (or attributes) and their values as soon as the progress in research proves it necessary. Each data card in the db-Builder contains a field for the annotated specimen of expression (an utterance chosen from a corpus) and a field with a list of attributes and values from which the linguists choose the relevant values for the sample they are describing.

The db-Builder module is complemented by that of the Tree Builder Assistant which allows the linguist to organise the chosen attributes and values in a tree structure. Any change in the feature tree of the Tree Assistant is featured in the database after the linguist confirms the changes. All samples are automatically collected in a contingency table. The synthetic table has the form of a chart for each sample described in the database.

For any linguistic description it is an important aid: it makes it possible to verify whether the same attribute and value were correctly chosen in different contexts and at different times of data description by the linguist. Such tables are processed by tools which provide statistical insight into the usefulness of attributes and suggest interactive arrangement of those attributes: it is possible (1) to check objects with duplicates, (2) to merge two or more attributes, (3) to show types of objects by attributes or by values, (4) to check the feature field input of each record and (5) to

display global or partial feature trees. Indeed, any automated check-up helps the linguist to verify the consistency of his work.

However, this approach is confronted by the serious problem of meta-data input; as a matter of fact, we claim that, just as is the case in natural sciences, *raw facts* in the domain of language do not exist as such and any description of linguistic reality relies on a felicitous set of meta-theoretical assumptions. At the stage of collecting and marking semantic data, the intuition of the linguist is unavoidable, yet fallible. One should be conscious that the process of annotating data relies heavily on the linguist's expertise in a given domain. Therefore, the choice of attribute value sets has to be discussed between younger and more experienced researchers. The best way to proceed is to begin with some attributes taken as axioms and to verify whether they are borne out by numerous data. The objective of man-machine interaction (consisting in handling lists of features) is to eliminate subjectivity (and variability of appreciation) towards the meaning of linguistic expressions.<sup>6</sup>

## 2.2. Data science methods (KDD or data mining)

Once a database has been collected, computer tools of knowledge discovery in databases (KDD) and, in particular, its data mining algorithms can reveal remarkably compound relations (usually “invisible” or “hidden”) from a very simple tabular representation of gathered data. KDD technology thus makes it possible (a) to transform tabular representations (or charts) into lattices (which are more powerful than trees because they allow multi-base inheritance), (b) to apply approximation techniques allowing reasoning with uncertain data and (c) to provide hierarchical analyses reflecting the mutual dependencies of data in the system. For this reason, dynamically built semantic maps (sets of words with attributes arranged by opposition relationships) and semantic lattices (sets of words with attributes arranged by entailment relationships), among others, are used during the description research.

---

<sup>6</sup> As an example, the computer-aided research with Semana led us to call into question the contemporary theory of gender in Polish grammar. We carried out a first trial of describing Polish gender morphology using the generally accepted theory of 5 (up to 9) gender values. In view of the inconsistent results we obtained, we had to put forward another theory keeping apart the *gender* category (with its three values: *masculine*, *feminine* and *neuter*) from the *animacy* category (*inanimate*, *animate-non-human* and *animate-human*). Thus, we ruled out the so-called “male-personal” gender value (*rodzaj męsko-osobowy*) and so were able to give account of utterances (contradicting the official theory) in which the verb bears the so-called “male-personal” ending whereas the subject of the utterance is composed of two or more nouns, none of which is at the same time masculine and “personal” (or human), e.g. *Dziewczyna i pies wychodzili codziennie o 5-ej.* (*The girl and the dog used to go out every day at 5 o'clock*). Cf. Włodarczyk H. 2010, 2018b.

Let us enumerate briefly *Semana's* symbolic analysers:

1. Formal Concept Analyser (FCA): a technique based on Lattice theory (Wille 1982); various functions for the analysis and processing of “Formal Concept Contexts” (single-valued tables)
2. Rough Set Analyser (RSA): a technique using approximation membership functions (Pawlak 1991); various functions for the analysis and processing of “information systems” (multi-valued tables)
3. Rough Formal Concept Analyser (RFCA): a combination of FCA with RSA; functions especially useful for searching similarities in formal concept contexts
4. Rough Decision Logic Analyser (RDLA): a combination of (a) Rough Set Analysis (RSA) and (b) Decision Logic Analysis (DLA) – a technique originating in Expert Systems technology; it can be viewed as a rough rule builder.

The statistical analyser consists of the Factor Correspondence Analysis (FCA) coupled with the Hierarchical Ascending Classification (HAC) in compliance with programs written by Jean-Paul Benzécri and co-workers in the 1970s at Paris University. Multi-valued tables containing symbolic values are converted into one-valued tables, called contingency tables. In turn, one-valued tables may be converted into Burt's tables (tables of co-occurrences). These tables are particularly useful for studying the dependence and clustering of attributes. The report gives the *eigenvalues* (inertia of the axes), the projections of each object and attributes onto the first 4 axes, and the contribution of each axis to the definition of each point and the contribution of each point to the definition of the axes. Projections in planes [1,2] and [1,3] are proposed by default, but any other plane may be represented. The HAC is also displayed, and classes may be coloured to help visualisation.

The meaning conveyed by natural languages is defined as a function from signs (in fact, from their schematic representations) into individualised ontologies. A formal cognitive description aims at giving an ontological account of semantic categories by treating their definitions in the form of finite sets of precisely defined *feature structures*. We first looked for building sets of abstract structures that we presumed would be useful for interpreting the categories of different languages we were studying, and then we chose a subset relevant to the study of a given language.<sup>7</sup> The use of computer tools for collecting databases and extracting knowledge out of them opens up a new era in linguistic research, making it possible for researchers (situated all over the world) to exchange data and discuss problems on a formalised basis.

---

<sup>7</sup> As an example, let us quote the description of Polish verbal aspect (Włodarczyk A., Włodarczyk H. 2006) with attributes and values extracted from general aspect ontology (Włodarczyk A. 2003).

Interactive research made it possible to find new and more adequate theoretical solutions for some fundamental linguistic problems such as (a) the construction of base and extended utterances,<sup>8</sup> (b) the construction of semantic components of situations (“states of affairs”, cf. Włodarczyk A., 2008) and (c) major grammatical categories<sup>9</sup> (e.g. aspect, type, movement verbs, modality, etc.).

### 3. The formal concept and the theory of signs

The mathematical definition of a **concept**, known as *formal concept*, is a dual pair of extension and intension within a given context of assignments of features (“intents”) to objects (“extents”). This definition, given by Rudolf Wille (1982), greatly inspired the data-mining community. Applying this definition to linguistic domains makes it possible, first of all, to treat symbols in the same way as “real world” objects, i.e. as both symbolic objects and ordinary objects, all of them considered as extents of formal concepts within their own systems (in FCA terms: “contexts”). In classic semiotics, concepts are either “signified” parts of signs which are “inseparable” from their “signifier” parts (Ferdinand de Saussure) or “interpretants” (a kind of link or function) of symbols with respect to objects (Charles S. Peirce). However, within the data-mining framework, studies of signs as objects are rare. Peircean semiotics was logically reconstructed by Uta Priss (2017) who participated in CELTA conferences twice. It is worth noting that this author published quite recently her add-on<sup>10</sup> to that theory (Priss 2020).

In the present approach, sign-objects are internalised as formal objects of the formal concepts called **semions** and real objects (things) are internalised as formal objects of the formal concepts called **noemata**. Thus, we came to the conclusion that the conception of linguistic sign should be developed in order to include it in the complex system of utterance meaning representation. This led to a semiotic rectangle<sup>11</sup> (Fig. 4) in place of the triangle put forward by Charles S. Peirce. Moreover, it is important to note that it seems quite possible that man thinks in both an

<sup>8</sup> As an example, note that our approach makes it possible to explain the choice between the so called “orthotonic” and “enclitic” forms of the Polish 1st person singular pronoun in the dative case *mnie* or *mi*, which is a difficult problem even for native speakers. The MIC theory brought to light the pragmatic conditions underlying the choice of these forms (Włodarczyk H. 2018a). But more generally, several other authors demonstrated that the speakers of other European languages such as Russian, French, Greek and Latin (cf. Włodarczyk A., Włodarczyk H. 2013) also need the meta-informative statuses for using either “orthotonic” or “enclitic” variants of personal pronoun forms.

<sup>9</sup> References to publications on these topics can be accessed online at <http://celta.paris-sorbonne.fr/DG-Biblio.html>.

<sup>10</sup> As a result of this recasting, Uta Priss obtained a sort of integration of the Peircean triangle with our semiotic square (cf. Włodarczyk A. 2017).

<sup>11</sup> It is impossible for us to present this problem in more detail here. Hopefully, an exhaustive presentation will be soon published in the Journal “Studies in Logic, Grammar and Rhetoric”.



internal conceptual code and a natural language defined in this completely new setting as a **quasi-autonomous semiotic system**.

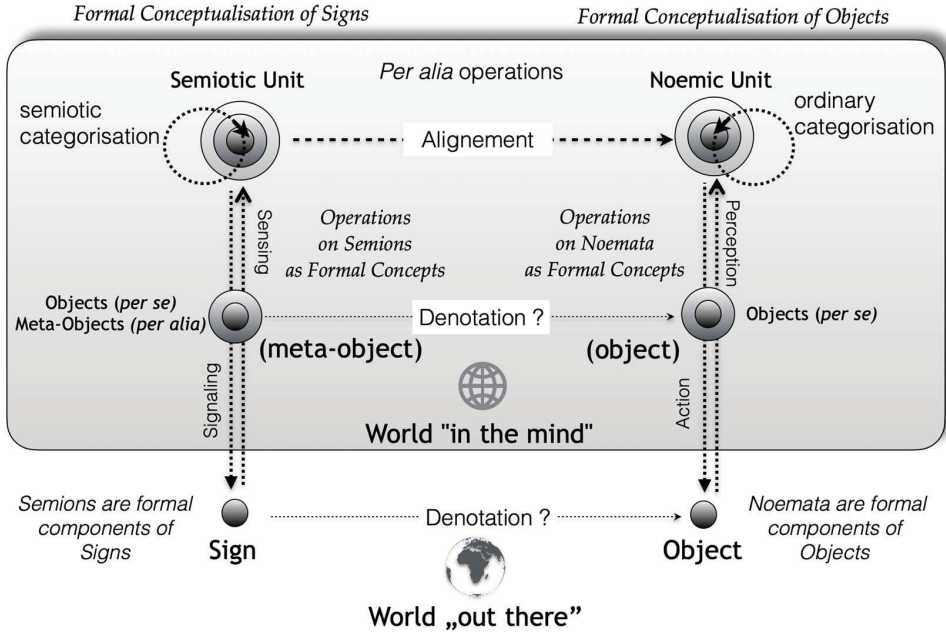


Fig. 4. Formal Semiotic and ordinary concepts as core components of information

As an example, let us compare (a) the linguistic formal object of the **semion** {{{“Mädchen”}}, {Part-of-Speech: noun, Gender: neuter}} in which “Mädchen” represents the internalised trace of the German word with (b) the ordinary formal object of the possibly universal **noema** {[girl]}, {Entity: human, Sex: female}} in which [girl] represents the internalised trace of the formal object of a thing, both concepts being represented in the speakers’ mind.

It turns out that, roughly speaking, in (1) the Saussurian theory of sign, the “signifier” is an “extent” and the “signified” is an “intent” of a semion and that (2) the Peircean “interpretant” should probably be seen as a function, not as a “concept”. In some other settings of the Peircean theory where the “interpretant” is interpreted as a “concept”, it should be split into two. Clearly, from a more general point of view, the problem of meaning is much more than an interpretation function from expression (form) to representation (content). Most probably it is a rather complex set of alignements (a) between semions and noemata, hence – a fortiori – between semiotic and noemic units, (b) between internal conceptual structures of traces of

physical expression units and of their cognitive counterparts (between realisations of *verb valence schemata* and instantiations of *cognitive protoforms*), on the one hand, and (c) between all kinds of ortho-, para- and meta-information, etc., on the other hand. In addition to this and perhaps most of all, it is the result of multiple composition relationships within and between various language unit layers.

#### 4. Modelling linguistic communication as a distributed conceptual system

The classical hierarchical representation of linguistic utterances needs to be enhanced with multi-dimensional representations, more suitable for building meaning within intelligent (multi-processor) distributed systems. *Distributed Grammar* was designed as a multi-dimensional approach to the modelling of natural language expressions. This analytical view emerged as the result of an investigation into syntactic structures and, especially, after it became clear that language reflects both semantic and pragmatic components of the sense of utterances. *Distributed Grammar* is therefore an integrated framework for our semantic and pragmatic theories of communicated information (*Associative Semantics* (AS) and *Meta-Informative Centering* (MIC) theory, respectively). This integration can be achieved using two kinds of downward contextual dependence (expansion) of meaning: (1) external – *grounding* which is applied in order to determine the informative **truth value** and (2) internal – *refinement* which in some cases is necessary for determining the meta-informative **aboutness** (including predication) and the meta-informative **status** of utterances treated as *old* or *new* information. Since, in *Distributed Grammar*, refinement also crosses some other spaces of discourse analysis such as the communicative space (“backward/forward looking centred” units or, more classically, anaphors/cataphors), the epistemic space (“known/unknown” and other kinds of meta-informative modalities such as belief, possibility etc.), it can be considered therefore that both grounding and refinement play the role of **liaison (glue)** between the pragmatic and semantic components of grammar.

This theoretical background makes it possible to apprehend, at one and the same time, **expression** (communication in all its aspects: verbal, visual but also using other channels and supports) and **perception** (cognitive experiencing of states and actions) which are simply extreme examples of using ingredients from a rich realm of meaningful elements of the world, starting with **symbols**, going through all kinds of other signs (indexes, icons, signals etc.) and ending up with **things** (objects and facts). ‘Symbols’ stand for linguistic signs and ‘things’ cover everything possible (no matter whether it is real or imaginary).

Because of the dual role of signs, the question of the dynamics of *semiosis* arises, which consists not only in its constant evolution, but also, and above all, in the parallel processing of various structures representing meanings. Introducing into linguistics the thesis about parallel structure processing made it possible to develop the *distributed grammar* program, according to which the primary role in the building / understanding of linguistic expressions is played by activating partial information **scattered** across various knowledge modules. This fact is closely linked to the interpretation of lexical-syntactic schemes called “verb valences”, to which, additionally, grammatical recombination rules are applied in order to convert them into parallel (often correlated) structures of semantic representations with different meta-informative perspectives (Włodarczyk A., Włodarczyk H. 2019).

#### 4.1. Understanding during the process of communication

To what extent do people understand each other in a conversation? In order to answer this question, we need to find out which information contained in an utterance is activated and processed by each participant, and this depends on the kind of relationship between signs and ordinary objects (be they real or imaginary). Two different levels are usually taken into account: (1) that of denotation as a direct relationship – one might say – the effect of naming objects and (2) that of representation – as an indirect relationship through the concept (meaning). While the first case is rather simple and belongs to the domain of logic (cf. Alfred Tarski), the latter is mainly dealt with by semantics and the philosophy of representation (cf. Charles S. Peirce).

In our view (Fig. 5), the answer to the question of the quality or degree of agreement in the communication process (i.e., with the participation of usually more than one epistemic agent) concerns the equivalence of representation resulting from the imaginary **unification** of the speaker’s intention (representation A) and the hearer’s comprehension (representation B).

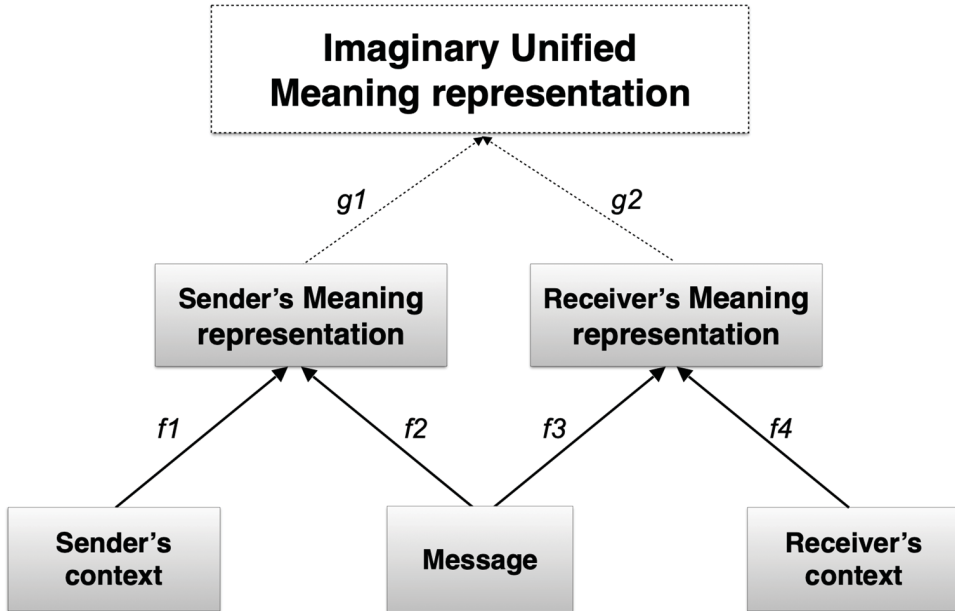


Fig. 5. The unification of speaker's and hearer's representations in an abstract distributed network

Thus, in order to give a clear answer to this question, it is necessary to define individual profiles of interlocutors, their knowledge, beliefs and conditions regarding the speech act, and this seems to be determined by the so-called **context dependence** of the content of utterances. For example, if the speaker were to use intentionally a partial utterance, the listener would have to undertake the extension of that content. However, in order to build models of speech acts, the researcher should put himself in an objective position with respect both to the speaker and the hearer. Thus, there is no doubt that such an approach is integrative in contrast to other alternative points of view, i.e., those which take into account only one participant of the communication, in other words consider only the *input* or the *output*.

#### 4.2. Layers of information in natural language

Linguists and computer scientists generally consider syntax along with semantics and pragmatics as different aspects of information. In our view, ontology is closer to semantics and pragmatics than to syntax. In fact, syntax concerns the material **form** of information belonging to all these three domains. The form of

natural language utterances includes prosody, phonology, morphology **and syntax** whereas their content consists of at least three layers of information within the identificational dimension: (1) the ortho-informative (“properly” or “literally” semantic) configuration in parallel with (2) the para-informative contrastive *identification* and (3) the meta-informative reflexive identification within the processes related to *aboutness* (*predication*, *topicalization* and *focalisation*). Elements which have been *identified* by contrast to others in the para-informative layer are *configured* in the ortho-informative one. In the meta-informative layer, elements of para- and ortho-information are in turn selected and ordered depending (a) on the speaker’s centres of attention and also, as far as shared attention is concerned, on the hearer’s centres of attention, as well as (b) on the specific means the speakers have at their disposal in the language they speak. Table 1 summarises the operations characteristic of each information layer.

Table 1. Intra-layer operations on information within utterance content

Information layers	Intra-layer operation	Result
<i>Meta-information</i>	Centering	Attention-driven chunks of information
<i>Ortho-information</i>	Configuring	Literal information content: situations & participants
<i>Para-information</i>	Contrasting	Components of semantic situations

#### 4.2.1. Para-information

The para-information layer is the space of such concepts which are overtly expressed in linguistic utterances but are established with regard not only to themselves but to other *similar* or even by contrast *opposed* concepts which are **not** expressed in the ortho-information of the given utterance, remaining implicit or covert, though they have to be taken into account in the process of interpretation.

Although identification is always present in the meta-informative layer, it is rarely expressed in a manner other than by naming an individual or a relationship. Morphemes such as “also”, “even” and “only, solely, merely” are used as *relative* identifiers which may be explicitly expressed in natural language utterances (Fig. 6). Para-information concerns therefore the identification of concepts (among them those of entities and situations) when they emerge in the speakers’ mind not only in *reflexive* terms (in relation to themselves) but also in a *relative* way as a kind of ‘aliqueness’ (as compared to other members of the same group of beings, class of figures, location in space and time etc.). It seems interesting that logicians who choose to work on reasoning with identity give the identity relationship a separate status among all other relationships.

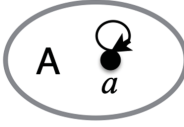
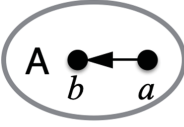
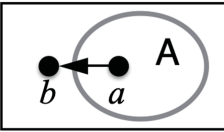
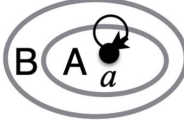
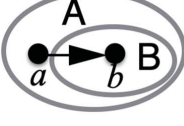
	<i>Identity</i>	<i>Alikeness</i>	
<b>complexity</b>	<b>Reflexive Identity</b>	<b>Relative Identity</b>	
		<b>Inclusive</b>	<b>Exclusive</b>
simple	$\emptyset a$  <i>Indication</i>	<i>also a</i>  <i>Comparison</i>	<i>only a</i>  <i>Exclusion</i>
	embedded	<i>just a</i>  <i>Insistence</i>	<i>even a</i>  <i>Concession</i>

Fig. 6. Identity expresses meta-information and alikeness expresses para-information

Note that adverbs, particles and deictic pronouns constitute the language expression of this identifying selection, as understood in a broad sense, in which equivalence and complementarity also participate (as distinctive semantic criteria).

#### 4.2.2. Ortho-information

The ortho-information layer is determined by an encapsulation/activation function within a certain information frame (a configured network of information) concerning primarily the roles of participants of the spatio-temporally anchored situation spoken-about in an utterance. In the ortho-informative layer, information about roles played by participants and spatio-temporal anchors is configured into situation frames. Conceptual schemes are mental representation frames through which it is possible to pick up and identify situations of the outer world (Minsky 1974). In the process of linguistic communication, conceptual schemes from the speaker's ontological knowledge are matched with **valence schemes**, sort of ready-made linguistic frames rendering it possible to express oneself very briefly. Native speakers acquire in their linguistic *milieu* knowledge about "verb valences" together with the roles played by participants when taking part in the situation expressed by the verb.

In human languages, there is no direct access to ortho-information in an utterance: only literal meaning is explicit whereas the cognitive content is barely alluded to and has to be reconstructed in the process of semantic interpretation. The semantic content of linguistic categories is distinct from their cognitive content, for instance, *gender* is not *sex*, nor *tense* in grammar is *time* in human experience, nor does grammatical *number* match directly the cognitive notion of *quantity*.

### 4.2.3. Meta-information

The meta-information layer is the one of attentional partitioning of ortho-information into 1st-degree meta-information and 2nd-degree meta-information for the purpose of discriminating central and peripheral parts of it. The first level of meta-information concerns that of base utterances in which the speaker's centres of attention are expressed by the subject (global) and object (local). The subject is this attention-driven phrase (ADP) which refers to the speaker's global centre of attention (CA). According to Klaus Oberauer (2003), a neurologist of attention, no more than two centres of attention (one global and one local) can be active at once; therefore, only the subject and direct object are attention-driven phrases. Other nominal phrases occurring in utterances directly refer to ortho-information. The verb and its complements are what is being said about the subject; they are the constituent parts of what we call *predicate*. The predicate reflects the meta-informative clustering of representation. Consequently, predication in natural language is not a semantic notion but a pragmatic one, since meta-information depends on the speaker's point of view, and more precisely, their own centres of attention presumably shared with the hearer. Due to the sequential order of linguistic discourse as a product of mental operations dealing with semantic situations which are probably both incremental and parallel, the **linguistic predicate**, *subject* and *direct object* are of pragmatic nature (Włodarczyk A., Włodarczyk H. 2016); therefore, these entities cannot be defined properly by reference to the **logical** notion of predicate and its arguments which belong to semantics (Włodarczyk A., Włodarczyk H. 2019).

The second level of meta-information concerns extended utterances in which the global ADP is the topic and the local one the focus. The most important property of topic and focus is their respectively old and new status of communicated information which is in contrast with the rest of the utterance. As a matter of fact, although this might seem at first glance contradictory with the definition of "information", communication in natural languages is based on the alternation of old and new. Moreover, the old or new status is not an inherent property of ortho-information but is based on the way in which it is communicated. The speaker is free to introduce some chunks of information either with a new or old meta-informative

status: this is a major argumentative device in the strategy selected to enrich or manipulate the hearer's knowledge.

Base and extended utterances are defined as pragmatic units of discourse in contrast to simple and complex sentences understood as syntactic units. As a pragmatic unit, each utterance contains at least one attention-driven phrase (ADP) referring to a centre of attention (CA). The ADP may have either the same or a different meta-informative status (old or new) from the rest of the utterance (the statement). In a base utterance, there is no contrast between the status of the global ADP and that of the statement: it is either "all new" or "all old". On the other hand, the ADPs of extended utterances contrast with the statement. The topic bearing an old meta-informative status is in contrast with the new comment, the focus of new meta-informative status is in contrast with the old background. When a constituent of a base utterance (be it an ADP or another NP) undergoes topicalisation or focalisation, it is treated as an extension of the base utterance which, in many languages, is moved to the left or right periphery of the base utterance. In the so-called free word order languages (e.g. Polish, Russian), topic and focus are superimposed on other NPs mostly by means of prosody and word order.

Concepts coined for the description of meta-information have been elaborated using algorithms from the Formal Concept Analysis (Wille 1982) and Rough Set Theory as implemented in the *Semana* platform. It is worth noting that only two features (Pawlak 1981, 1991) – the *old* or *new* status, and the global or local property of attention-driven phrases – suffice to show the variety of attention-driven phrases in the form of a lattice (Fig. 7).

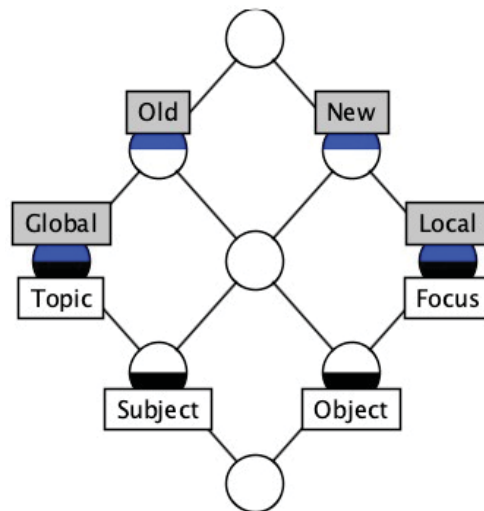


Fig. 7. Lattice revealing the interdependencies of attention-driven phrases



Base utterances are generated using valence schemes. In order to change the meta-informative structure of an utterance, languages make use of *recombination* rules, with some operating within base utterances and other making it possible to build extended utterances. Verb voice (active/passive) belongs to the meta-informative grammatical means used in recombination, cf. utterances (1) and (2).

- (1) Brutus killed Caesar.
- (2) Caesar was killed by Brutus.

With the passive voice, the distinction of salience (global/local) is changed without changing the ortho-informative content of the utterance, i.e. the roles of *killer* and *killed* are played by the same participants, [Brutus] and [Caesar], respectively. The recombination of ADP by topicalisation or focalisation makes it possible to build *extended* utterances: the most universal means are prosody and word order, but particles and syntactic constructions are also commonly used (Włodarczyk A., Włodarczyk H. 2013, 2019).

#### 4.2.4. Integration of information layers

Integration of information belonging to the three different layers is enabled by the fact that participants and spatio-temporal anchors of the situation are identified as referring to the same entities. As an example, we represent on Fig. 8 the three layers of information expressed in the utterance (1).

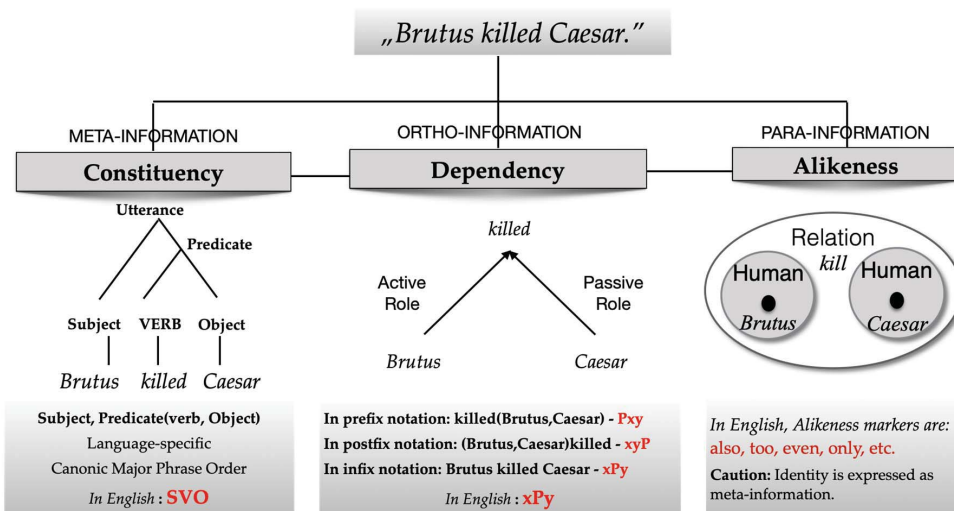


Fig. 8. Coincident Information Structures within the utterance “Brutus killed Caesar”

Importantly, what is immediately conceived and perceived during the synthesis and analysis of utterances is their meta-informative structure, under the cover of which other layers of information remain implicit. Ortho- and para-information need more processing effort for producing and understanding linguistic utterances.<sup>12</sup> In order to complete successfully the comprehension process, the interpretation of an utterance has to get access to the ortho-informative layer of semantic representations.

## 5. Conclusion

As shown above, the interactive method applied to language data provides a good reason for a real shift in approaching natural language objects that might result in a new paradigm in general linguistics<sup>13</sup> which we provisionally call “Conceptual Linguistics”. We use the qualifier “conceptual” because it is based on formally defined concepts (Wille 1982), replacing the classic notion of sign – understood interpretively ‘purely formally’ as a symbol by that of a **semion** (i.e.: a *semiotic concept*), since signs have no meaning outside the cognitive systems which conceive and process them. For this reason, we use semions seen as signs dually defined as couples of *extent* (their “form”) and *intent* (their “content”). As a result, from the point of view of logic, the emerging theory of conceptual linguistic semantics will be neither extensive nor intensive, but dually interpreted as consisting of both extension and intension.

Research in the spirit of *Conceptual Linguistics* consists in modelling speech-acts by distinguishing *semiotic-semantic* operations (which belong to the domain of linguists) from *ontological-semantic* operations (which belong to the domain of psychologists, sociologists, cultural scholars and logicians) bearing in mind that these semantics are correlated in a parallel fashion. The task of *modelling natural language* system operations belongs to the domain of computer simulation specialists being aware, however, that the engineering knowledge, which is today available, will hardly quickly suffice for their Real-World applications.

From the point of view of language philosophy, *Conceptual Linguistics* is supposed to be, in a sense, a compromise between two extreme views on the key problems of today’s linguistics, between the universalistic and relativistic views (Włodarczyk H. 2018b). This approach is made possible primarily by treating signs as objects that play a dual role in the process of communicating content, i.e.: not

---

<sup>12</sup> At a workshop organised by Wioletta Miskiewicz in Paris in 2008, Jan Woleński commented on Włodarczyk’s approach, suggesting that “it is quite possible that, in natural language utterances, truth lies much deeper than the logicians suppose”.

<sup>13</sup> And more generally, in semiotics.

only as objects *per alia*, but also as objects *per se*. This view can be generalised to cover all the **multi-modal systems** (systems of signs), on the one hand, and all the **mental systems** (including multi-sensory ones) – on the other. It should also be pointed out that the semantics of all these systems are based on Formal Concepts (they are, as a rule, mathematically definable), otherwise it would be difficult to imagine on what principles, other than mathematical, biological systems of such high complexity could ever function.

## Bibliography

**Nota bene:** References to other publications can be found at these website pages:

- (1) Interactive Linguistics <http://celta.paris-sorbonne.fr/anasem/papers/>.
- (2) Distributed Grammar <http://celta.paris-sorbonne.fr/DG-Biblio.html>.

- Benzécri J.-P., 1984, *L'analyse des données*, 4ème éd., vol. 1: *La Taxinomie*, vol. 2: *L'Analyse des correspondances*, Paris: Dunod.
- Fayyad U., Piatetsky-Shapiro G., Smyth P., 1996, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine* 17 (3): 37–54.
- Minsky M., 1974, *A Framework for Representing Knowledge*, MIT-AI Laboratory, Memo 306.
- Oberauer K., 2003, "Selective Attention to Elements in Working Memory", *Experimental Psychology* 50: 257–269.
- Pawlak Z., 1981, "Information Systems – Theoretical Foundations", *Information Systems* 6 (3): 205–218.
- Pawlak Z., 1991, *Rough Sets. Theoretical Aspects of Reasoning about Data*, Dordrecht: Kluwer Academic Publishers.
- Pogonowski J., 1981, *Tolerance Spaces with Applications to Linguistics*, Poznań: Adam Mickiewicz University Press.
- Pogonowski J., 1993, *Linguistic Oppositions*, Poznań: Adam Mickiewicz University Press.
- Priss U., 2017, "Semiotic-conceptual Analysis: A Proposal", *International Journal of General Systems* 46 (5), Special Issue on Concept Lattices and Their Applications: 569–585.
- Priss U., 2020, "A Preliminary Semiotic-Conceptual Analysis of a Learning Management System", in: M. Cristani, C. Toro, C. Zanni-Merk, R. J. Howlett, L. C. Jain (eds.), *Knowledge-Based and Intelligent Information & Engineering Systems*, Proceedings of the 24<sup>th</sup> International Conference KES 2020, Vol. 176, Elsevier, 3702–3709.
- Sauvet G., 2008, "Symbolic and Statistical Analyses of Meta-data Using the "Semana" Platform – A Bundle of Tools for the KDD Research", PPT presentation, CASK Sorbonne 2008 (Language Data Mining) International conference, Université Paris-Sorbonne – Paris 4.
- Stacewicz P., Włodarczyk A., 2010, "Modeling in the Context of Computer Science – A Methodological Approach", *Studies in Logic, Grammar and Rhetoric*, special issue: Philosophical, Trends in the 17<sup>th</sup> Century from the Modern Perspective, H. Świączkowska (ed.), 20 (33): 155–179.
- Stacewicz P., Włodarczyk A., 2020, "To Know We Need to Share – Information in the Context of Interactive Acquisition of Knowledge", in: M. Cristani, C. Toro, C. Zanni-Merk, R. J. Howlett, L. C. Jain (eds.), *Knowledge-Based and Intelligent Information & Engineering Systems*, Proceedings of the 24<sup>th</sup> International Conference KES 2020, Vol. 176, Elsevier: 3810–3819, <http://kes2020.kesinternational.org/cms/userfiles/is23.pdf>.
- Wille R., 1982, "Restructuring Lattice Theory: An Approach based on Hierarchies of Concepts", in: I. Rival (ed.), *Ordered Sets*, Dordrecht–Boston: Reidel, 445–470. Reprinted in: S. Ferre,

- S. Rudolph (eds.), *Formal Concept Analysis*, 7<sup>th</sup> International Conference, ICFCA 2009 Proceedings, Heidelberg: Springer, 314–339.
- Wittgenstein L., 1953, *Philosophical Investigations*, G. E. M. Anscombe, R. Rhees (eds.), G. E. M. Anscombe (trans.), Oxford: Blackwell.
- Włodarczyk A., 2003, “Les Cadres des situations sémantiques”, *Études Cognitives / Studia Kognitywne* 5: 35–51 (English translation: (2013) “Frames of Semantic Situations”, in: A. Włodarczyk, H. Włodarczyk (eds), *Meta-informative Centering in Utterances – Between Semantics and Pragmatics*, Amsterdam: John Benjamins Publishing Co., 21–40.
- Włodarczyk A., 2007, ハリ・ソルホンヌ大学 理論・応用言語学研究所 (CELTA) — CASK (Computer-aided Acquisition of Semantic Knowledge)プロジェクト — (paper in Japanese), in: *Japanese Linguistics*, Vol. 21, Tokyo: The National Institute for Japanese Language, <http://celta.paris-sorbonne.fr/anasem/papers/miscelanea/CELTA-CASK-AW-E.pdf>.
- Włodarczyk A., 2008, “Roles and Anchors of Semantic Situations”, *Études Cognitives / Studia Kognitywne* 8: 53–70.
- Włodarczyk A., 2015, “Informatyka szansą na rozwój naukowej lingwistyki” (“Computer Science as an Opportunity for the Development of Scientific Linguistics”), in: P. Stacewicz (ed.), *Od informatyki i jej zastosowań do światopoglądu informatycznego*, Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 117–132.
- Włodarczyk A., 2017, “Pojęcie a znak w świetle systemów informacyjnych” (“Concept and Sign in the Light of Information Systems”), *Filozofia w informatyce*, 45. spotkanie, Centrum Kopernika Badań Interdyscyplinarnych, UPJPII, Kraków, <https://filozofiainformatyki.wordpress.com/>.
- Włodarczyk A., Włodarczyk H., 2006, “Semantic Structures of Aspect, a Cognitive Approach”, in: I. Bobrowski, K. Kowalik (eds.), *Od fonemu do tekstu. Prace dedykowane Romanowi Laskowskiemu*, Kraków: Instytut Języka Polskiego PAN, Lexis, 389–408.
- Włodarczyk A., Włodarczyk H. (eds.), 2013, *Meta-Informative Centering in Utterances – Between Semantics and Pragmatics*, Companion Series in Linguistics N°143, Amsterdam: John Benjamins.
- Włodarczyk A., Włodarczyk H., 2016, “O pragmatycznej naturze predykcji (czyli o metainformacji w orzekaniu językowym)”, *Poradnik Językowy* 8: 7–21.
- Włodarczyk A., Włodarczyk H., 2017, “Subjecthood and Topicality are both Pragmatic Issues”, in: Y. Harada, S. Shudo, M. Takekuro (eds.), *Papers on and around the Linguistics of BA*, Tokyo: Waseda University, 1–10.
- Włodarczyk A., Włodarczyk H., 2019, “Qu’est-ce au juste que la prédication ?”, *Bulletin de la Société de Linguistique de Paris* 64 (1): 1–54.
- Włodarczyk H., 2010, “Lingwistyka na polonistyce krajowej i zagranicznej w dobie filozofii informatyczno-logicznej” (“Polish Linguistics in Poland and Abroad in the Age of Logico-computational Philosophy”), *LingVaria* 1 (7): 65–79.
- Włodarczyk H., 2018a, “Mnie czy mi? O użyciu zaimka pierwszej osoby w celowniku” (“Mnie or mi (me)? On the Use of the 1<sup>st</sup> Person Pronoun in the Dative Case”), *Poradnik Językowy* 9: 64–80.
- Włodarczyk H., 2018b, “O potrzebie wiedzy o języku polskim kompatybilnej z wiedzą o innych językach narodowych” (“On the Need of Concepts of Polish Grammar which Would Be Compatible with Those of Other National Languages”), in: A. Achтели, K. Graboń (eds.), *Polonistyka na początku XXI wieku. Diagnozy. Koncepcje. Perspektywy*, t. 5, Katowice: Wydawnictwo UŚ, 205–218.